Inferring Speaker Affect in Spoken Natural Language Communication

Dissertation advisor

Professor Stuart M. Shieber

Author

Heather Roberta Pon-Barry

Inferring Speaker Affect in Spoken Natural Language Communication

Abstract

The field of spoken language processing is concerned with creating computer programs that can understand human speech and produce human-like speech. Regarding the problem of understanding human speech, there is currently growing interest in moving beyond speech recognition (the task of transcribing the words in an audio stream) and towards *machine listening*—interpreting the full spectrum of information in an audio stream. One part of machine listening, the problem that this thesis focuses on, is the task of using information in the speech signal to infer a person's emotional or mental state.

In this dissertation, our approach is to assess the utility of prosody, or manner of speaking, in classifying speaker affect. *Prosody* refers to the acoustic features of natural speech: rhythm, stress, intonation, and energy. *Affect* refers to a person's emotions and attitudes such as happiness, frustration, or uncertainty. We focus on one specific dimension of affect: level of certainty. Our goal is to automatically infer whether a person is confident or uncertain based on the prosody of his or her speech. Potential applications include conversational dialogue systems (e.g., in educational technology) and voice search (e.g., smartphone personal assistants).

There are three main contributions of this thesis. The first contribution is a method for eliciting uncertain speech that binds a speaker's uncertainty to a single phrase within the

larger utterance, allowing us to compare the utility of contextually-based prosodic features. Second, we devise a technique for computing prosodic features from utterance segments that both improves uncertainty classification and can be used to determine which phrase a speaker is uncertain about. The level of certainty classifier achieves an accuracy of 75%. Third, we examine the differences between perceived, self-reported, and internal level of certainty, concluding that perceived certainty is aligned with internal certainty for some but not all speakers and that self-reports are a good proxy for internal certainty.

Contents

List of Figures

List of Tables

Acknowledgments

Producing this thesis would not have been possible without the support of several colleagues, friends, and family. First, I wish to thank my advisor Stuart Shieber, for thinking critically, listening patiently, and always believing in me. With his guidance, I have grown into an independent researcher, improved my communication skills, and added to my repertoire several typesetting tricks. Next, I am grateful to Barbara Grosz and Krzysztof Gajos, who have given me invaluable feedback about my research and have helped me see the bigger picture.

Thank you to David Parkes, Patrick Wolfe, Radhika Nagpal, Margo Seltzer, Ryan Adams, Yiling Chen, and Avi Pfeffer for your guidance and support, to Cassandra Extavour for aiding my professional development, to Abeer Alwan for helping shaping my early research ideas, and to Stanley Peters and Fuliang Weng for the encouragement and mentorship that led me to pursue a Ph.D.

Elif Yamangil, thank you for being a superb officemate, colleague, and friend. Ece Kamar, Anna Huang, Stacy Wong, Elaine Angelino, Karl Schultz, Jung Hong, Dan Rudoy, Sam Wiseman, members of AIRG, and members of HGWISE, I am grateful for your friendship, your feedback, your patience, your explanations, and your generosity.

I have had the opportunity to work with several talented undergraduates. Nick Longenbaugh, Vanessa Tan, Spencer de Mars, and Enrique Henestroza — thank you for your diligence and enthusiasm.

I am grateful for my friends in California, especially Elaine Chao and Jess Dang, for providing levity and adding the phrase "natural language processing" to your lexicon. I am grateful for the friends I have made in Boston, especially on the soccer field. The spirited and steadfast members of the Caddis Fliers, Tailgators, and Team Long Gone have played

an integral role in maintaining my sanity during these past years at Harvard.

I also wish to thank my extended family on the East Coast, you have made me feel at home in New England. Louise, thank you for your endless compassion and support, especially during times of transition. And last but by no means least, mom and dad, you have given me an extraordinary amount of encouragement, advice, support, and love. I could not have made my way through graduate school without you.

For my family, friends, and mentors.

Chapter 1

Introduction

Speech interfaces are now a familiar part of our everyday lives. Yet, while most people can cite an instance where they interacted with a call-center dialogue system or a command-based smartphone application, few would argue that the experience was as natural or as efficient as conversing with another human. To make intelligent systems that can communicate with humans using natural language, we need to address two subproblems. First, such a system not only needs to recognize the words that a user utters, but also must interpret the meaning of the words and of the utterance in the context of any emotions or attitudes that the user conveys. Second, the system needs to adapt its output to the user, given this nuanced understanding of what the user is conveying. The focus in this thesis is on the first subproblem: recognizing a user's internal mental state.

This thesis examines the relationship between speaker affect and prosody. SPEAKER AFFECT includes emotions and attitudes such as happiness, frustration, and uncertainty. PROSODY refers to the acoustic features of natural speech: rhythm, stress, intonation, and energy. It is widely accepted that prosody conveys a layer of communicative meaning

beyond the linguistic meaning of the words that are spoken (Hirschberg, 2003). However, obtaining quantitative measures of such communicative meaning is a challenge: the perception of speaker affect in naturally-occurring speech is highly subjective; good agreement among human annotators is typically halfway between perfect unison and agreement by chance.

In this thesis, I focus on one specific dimension of speaker affect: uncertainty versus confidence (i.e., level of certainty). When people are conversing face-to-face, listeners are able to sense whether the speaker is certain or uncertain from a combination of contextual, visual, and auditory cues (Krahmer and Swerts, 2005). If we enable computers to sense a speaker's level of certainty, potential applications include conversational dialogue systems (Forbes-Riley and Litman, 2009), language learning systems (Alwan et al., 2007), voice search applications (Paek and Ju, 2008; Wang et al., 2011), and speech-based smartphone assistants. Automatically inferring a speaker's level of certainty can improve the naturalness of these speech interfaces by allowing the systems to adapt their behavior to the user based on his or her level of certainty.

Although humans can convey their level of certainty through audio and visual channels, this thesis focuses on the audio (the speaker's prosody) because in many potential applications, there is audio input but no visual input. On tasks ranging from detecting frustration (Ang et al., 2002) to detecting flirtation (Ranganath et al., 2009), prosody has been shown to convey information about a speaker's emotional and mental state (Lee and Narayanan, 2005) and about their social intentions. Our work builds upon this, as well as a small body of work on identifying prosodic cues to level of certainty (Krahmer and Swerts, 2005) and classifying a speaker's certainty (Liscombe et al., 2005). The intended application of

such work is for dialogue systems to respond appropriately to a speaker based on their level of certainty as exposed in their prosody, for example, by altering the content of system responses (Forbes-Riley et al., 2008) or by altering the emotional coloring of system responses (Acosta and Ward, 2009).

Our primary goal is to determine whether prosodic information from a spoken utterance can be used to determine how certain a speaker is. We argue that speech-based applications will benefit from knowing the speaker's level of certainty. The term "level of certainty" has multiple interpretations. It may refer to how certain a person sounds to a hearer, the *perceived* level of certainty. We want our system to hear whatever it is that humans hear. Not surprisingly, this is the definition that has been assumed in previous work on classifying level of certainty (Liscombe et al., 2005; Liscombe, 2007). However, in applications such as spoken tutoring systems (Litman and Silliman, 2004; Forbes-Riley et al., 2008) and second language learning systems (Alwan et al., 2007), we would like to know how certain speakers actually are — their *internal* level of certainty, in addition to how certain they are perceived to be.

Knowledge of internal level of certainty affects the inferences intelligent systems can make about the speaker's cognitive state, for example, whether the speaker has a misconception, makes a lucky guess, or might benefit from some encouragement. Getting a ground truth measurement of a speaker's internal level of certainty is a challenging problem that we address in this thesis. First, we ask speakers to rate their own level of certainty, we refer to this as the *self-reported* level of certainty. We also collect speech data using materials designed to control the speaker's actual, internal level of certainty. One novel contribution of this thesis is the collection and comparison of these different measures of uncertainty.

In the prior work on using prosody to classify level of certainty, no one has attempted to measure a person's internal level of certainty.

1.1 Related Work

The area of inferring a person's affect based on the prosodic information in their speech is an active and growing area of research. Some of the early work in this area was aimed at distinguishing positive and negative emotion (Lee and Narayanan, 2005; Litman and Forbes-Riley, 2006) and at detecting the emotions of annoyance and frustration (Ang et al., 2002). The latter had application in dialogue systems for call centers; if a caller is exasperated by the automated dialogue system, then it is in the call center's interest to transfer the caller to a live operator. These systems reported accuracies ranging from 71–77% when detecting annoyance and frustration (as a joint category) and using only prosodic features.

More recently, researchers have focused on topics related to speaker intent and social relationships in human-to-human conversation. Tepperman et al. (2006) build a model to detect sarcasm—specifically sarcastic versus neutral instances of the phrase, "yeah right" in the Switchboard corpus (Godfrey et al., 1992). Black et al. (2011) address the topic of detecting blame in conversations between spouses attending couples therapy with a counselor. Ranganath et al. (2012) look at detecting friendliness and flirtatiousness in speed dates. These studies all use a combination of prosodic and lexical features.

A central challenge in the area of affect classification is the lack of a clear gold standard. Typically, emotions and affect are manually annotated, and the notion of what constitutes "good" agreement varies depending on the type of affect being annotated. Further, there have not been standard data sets, as there are in some subareas of natural language process-

ing, such as parsing and machine translation. A step toward addressing these challenges has been taken by the European Union's HUMAINE Association (Schröder and Cowie, 2006), which attempts to create a community and put forth standards on methods and corpora related to emotions in human-computer interaction. The speech subarea of this community has organized challenge tasks in recent years, much like the organized shared tasks ("bake-offs") in the natural language processing community (e.g., SENSEVAL, CoNLL). They held the first Emotion challenge on positive vs. negative emotion detection while the second challenge focused on qualities of physical state: sleepiness and intoxication, which are relevant for car driver safety (Schuller et al., 2010, 2011).

Automatic detection and classification of affect has a wide range of potential applications. The applications fall into two broad categories: (1) systems intended for use by humans in human-human communication, and (2) systems meant to engage *with* humans, i.e., human-computer interfaces (Cowie et al., 2001). The first category spans areas in which automatic detection of affect can augment human judgement, for example, in detecting deceit, or in detecting flat affect, which is useful for diagnosing health disorders. The second category includes language learning systems (Alwan et al., 2007), spoken dialogue systems such as in-car dialogue systems (Pon-Barry et al., 2006b), voice search on smartphones (Paek and Ju, 2008), and tutorial dialogue systems (Litman and Silliman, 2004; Pon-Barry et al., 2006a), which we discuss further in section 1.2.

1.2 Uncertainty in Educational Technology

A goal in many tutorial dialogue systems is to emulate a skilled human teacher's ability to sense how well a student is grasping new material, and then tailor the lesson or discus-

sion accordingly (VanLehn et al., 2003; Pon-Barry et al., 2006a; Forbes-Riley and Litman, 2009). Knowing a user's internal and perceived levels of certainty allows such systems to assess cognitive state and speaker personality characteristics. Specifically, with estimates of internal and perceived certainty for a given utterance, intelligent systems can make inferences about the user's *self-awareness* and *transparency*.

We consider speakers to be SELF-AWARE if they feel certain when correct and feel uncertain when incorrect. The four possible combinations of correctness versus internal level of certainty are illustrated in Figure 1.1.

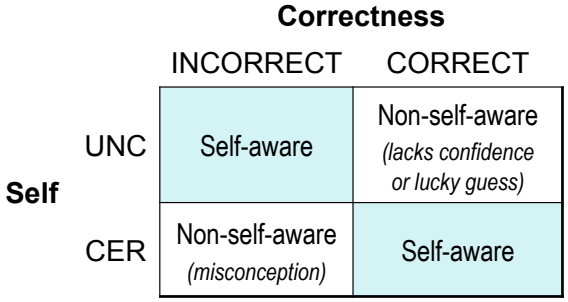

Figure 1.1: Self-awareness: speakers are self-aware if their internal level of certainty reflects the correctness of their answer.

Self-awareness is similar (though not identical) to the 'feeling of knowing' measure of Smith and Clark (1993). In conversational, question-answering settings, speakers systematically convey their feeling of knowing through both auditory and visual prosodic cues (Swerts and Krahmer, 2005).

In educational applications, if a student feels uncertain and is correct, it may be due to a *lack of confidence* or perhaps just a lucky guess. An intelligent system could ask a follow-up question to distinguish these two cases. Conversely, if a student feels certain and is incorrect, then it may be due to a *misconception*. Studies of learning in human tutorial dialogue suggest a strong connection between impasses (such as misconceptions)

and student learning, to the point of proposing that "cognitive disequilibrium" is a necessary precursor to deep learning (VanLehn et al., 2003; Craig et al., 2004).

The concept of speaker TRANSPARENCY is independent of an utterance's correctness. Speakers are transparent if they are perceived as certain when they feel certain and are perceived as uncertain when they feel uncertain. The four possible combinations of perceived versus internal level of certainty are illustrated in Figure 1.2.

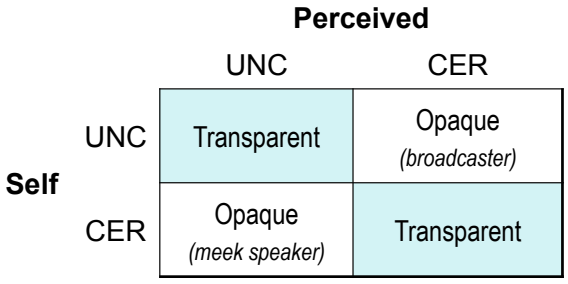

Figure 1.2: Transparency: speakers are transparent if their internal level of certainty reflects their perceived level of certainty.

A concept closely related to transparency is the 'feeling of another's knowing' (Brennan and Williams, 1995) — a listener's perception of a speaker's feeling of knowing (Smith and Clark, 1993). Recent work indicates that spoken tutorial dialogue systems can better predict student learning gains by monitoring the feeling of another's knowing than by monitoring only the correctness of the student's answers (Litman and Forbes-Riley, 2009). If an adaptive system determines what feedback to give a user based on a level of certainty model that is trained on perceived certainty ratings, then it will give inappropriate feedback to users who are *not* transparent.

We will discuss these concepts of self-awareness and transparency in the context of our Uncertainty Corpus in section 2.4.3.

1.3 Thesis Overview

Chapter 2 presents our motivation for creating the Uncertainty Corpus and describes a novel methodology for eliciting utterances of varying levels of certainty. This includes measuring level of certainty from both the hearer's perspective and the speaker's perspective: we collect level of certainty self-reports from the speakers and perceived level of certainty ratings from a panel of human judges. Further, in section 2.3, we present data collection materials that attempt to control a speaker's *internal* level of certainty. We find that self-reported certainty and internal certainty are highly correlated. This result is novel; the comparison of internal versus self-reported affect has not received significant attention in the affect detection community.

The Uncertainty Corpus is unique in that it contains sets of utterances that are lexically identical but differ in their level of certainty; thus any differences in prosody can be attributed to the speaker's level of certainty. Further, we control which words or phrases within an utterance are responsible for variations in the speaker's level of certainty.

Chapter 3 gives an overview of theories of intonational meaning and presents our prosodic analysis. We focus on using low-level prosodic features that are easily computed from the speech signal.

In chapter 4 we examine the use of prosody in modeling level of certainty from the hearer's perspective. We first present a basic machine learning model that uses prosodic information to classify utterances as certain, uncertain, or neutral. This model performs better than a trivial baseline model (choosing the most common class), corroborating results of prior work, but we also show for the first time that the prosody is crucial in achieving this performance by comparing to a substantive non-prosodic baseline.

In some applications, for instance, language learning and other tutorial systems, we have information as to which word or multi-word expression in an utterance is the LOCUS OF UNCERTAINTY, the probable source of uncertainty. We ask whether we can improve upon the basic model by taking advantage of this information. We show, in section 4.3, that the prosody of this locus of uncertainty and of its surrounding regions helps make better certainty classifications. Our best model reaches a classification accuracy of 75%. Conversely, in section 4.4 we show that our models can be used to make an informed guess, with an accuracy of over 90%, about which phrase a person is uncertain about when we do not have a priori knowledge of the locus of uncertainty.

In chapter 5 we discuss modeling level of certainty from the speaker's perspective. We look at whether simple machine learning models can classify self-reported level of certainty based on the prosodic features of the utterance. We also show that knowing the utterance's perceived level of certainty helps make more accurate predictions about the self-reported level of certainty.

Chapter 6 summarizes the conclusions of this thesis and describes areas of future work.

1.4 Contributions

In summary, the four main contributions of this dissertation are:

1. a methodology for collecting uncertainty data, plus an annotated corpus;

2. an examination of the differences between uncertainty from a hearer's point of view and from the speaker's point of view;

3. corroboration and extension of previous results in predicting perceived level of

certainty; and

4. a technique for computing prosodic features from utterance segments that both improves uncertainty classification and can be used to determine the locus of uncertainty.

Chapter 2

The Uncertainty Corpus

One central contribution of this thesis is a novel method for eliciting affective speech. The design of this data collection method enables us to analyze speaker level of certainty at both the utterance level and word level. The data collection method is motivated by four main criteria.

Criterion 1 For each speaker, we want to elicit utterances of varying levels of certainty.

Criterion 2 We want to isolate the words or phrases within an utterance that could cause the speaker to be uncertain.

Criterion 3 To ensure that differences in prosody are not due to the particular phonemes in the words or the number of words in the sentence, we want to collect utterances across speakers that are lexically similar.

Criterion 4 We want the corpus to contain multiple instances of the same word or phrase in different contexts.

In addition to collecting affective speech that meets these criteria, we want to ensure that our annotation process for generating level of certainty measures is as comprehensive as possible. To this end, we ask the speakers to self-report their own internal level of certainty. We also have external listeners (human annotators) judge each utterance's perceived level of certainty. The decision to measure both perceived and self-reported certainty is a key difference that distinguishes this work from prior work not only on inferring uncertainty but also on inferring speaker affect in general.

This chapter is organized as follows. Section 2.1 describes related speech data sets and their limitations. Section 2.2 presents the first phase of the data collection, for two domains: vocabulary, transportation. Section 2.3 describes the second phase: handwritten digits, with focus on controlling a speaker's internal level of certainty. Section 2.4 gives the logistical details of how the three kinds of uncertainty (perceived, self-reported, and internal) are measured. Lastly, section 2.5 contains descriptive statistics of the resulting Uncertainty Corpus.

2.1 Related Work

Research in the area of prosody-based affect detection has traditionally used corpora of spontaneous speech or corpora of acted speech (Ang et al., 2002; Lee and Narayanan, 2005; Liscombe et al., 2005; Tepperman et al., 2006). Although our ultimate goal is to create systems that can sense a speaker's affect in spontaneous speech, there are drawbacks to using the existing corpora of spontaneous speech. If the speech was originally collected for another purpose (as is the case in the tutorial dialogue speech (Litman and Silliman, 2004) examined by Liscombe et al. (2005)), there is no way to control the emotions and

affect of the speaker, and analysis is limited to perceived affect—there is no way to measure the speaker's internal affective state. A notable exception is recent work that measures both speaker and hearer perception of affect in the context of speed dates (Ranganath et al., 2012).

Acted emotional speech, on the other hand, attempts to produce speech with complete control of the speaker's affective state. However, the way that trained actors convey emotions may not be representative of the way that non-actors convey emotions.

In pilot experiments with both spontaneous speech and non-spontaneous speech, we found that the same set of prosodic features were significantly correlated with perceived level of certainty in both conditions. For this reason, the data collection method that we design is based on non-spontaneous read speech with isolated words that the speakers spontaneously select.

2.2 Vocabulary and Transportation Domains (Phase 1)

To collect speech data that meets the four criteria outlined at the start of the chapter, we designed a novel method for eliciting affective speech. The first phase of data collection resulted in utterances of varying levels of certainty in two topic domains: vocabulary and transportation. This section describes the speech elicitation method. Section 2.2.1 describes the materials, section 2.2.2 the participants, and section 2.2.3 the procedure. The Phase 1 data collection procedure has been described in previously published articles (Pon-Barry, 2008; Pon-Barry and Shieber, 2011).

2.2.1 Materials

The Phase 1 speech elicitation materials consist of a set of sentence templates—sentences that contain a blank along with multiple options for filling in the blank. The multiple options to choose among occur at the word or phrase level, while the rest of the sentence is fixed. By having participants read aloud a sentence template, we achieve consistency across utterances (criterion 3). However, we couldn't have participants just read a given sentence. Instead, to ensure varying levels of certainty (criterion 1), the participants are given multiple options of what to read and thus are forced to make a decision. Further, this allows us to isolate the phrases causing uncertainty (criterion 2).

Consider the two sentence templates below. The first example is in the domain of answering questions about using public transportation in Boston. The second example is in the domain of choosing vocabulary words to complete a sentence.

(Ex. 1) `Question: How can I get from Harvard to the Silver Line?`

`Answer: Take the red line to _____.`

`a. South Station`

`b. Downtown Crossing`

(Ex. 2) `Only the _____ workers in the office laughed at all of the manager's bad jokes.`

`a. pugnacious`

`b. craven`

`c. sycophantic`

`d. spoffish`

For the items in the transportation domain, we instruct the participants to imagine that the experimenter has recently moved to the Boston area and has stopped them on the street to ask for directions. This scenario is modeled after the Boston Directions Corpus task scenarios (Nakatani et al., 1995; Nakatani, 1997); we utilize this scenario to make the interaction feel less artificial and more like human-human conversation. For the items in the vocabulary domain, we simply instruct the participants to read the sentence aloud as if they are talking to the experimenter.

The method for presenting the sentence templates and multiple options for filling in the blank to the participants is described in section 2.2.3.

2.2.2 Participants

Twenty volunteers from the Harvard community participated in the Phase 1 data collection. All participants were native English speakers. There were 14 females and 6 males. The mean age was 22.35 (standard deviation 3.13).

2.2.3 Procedure

Participants interact both with the experimenter and with a computer interface throughout the speech elicitation process. Their speech is recorded through a head-mounted microphone.

Consider Example 1 above. In this example, the experimenter asks the participant, *How can I get from Harvard to the Silver Line?* The question is only spoken; it is not displayed visually. Without seeing the options for filling in the blank, the participants see the fixed part of the response, *Take the red line to* _____. They have unlimited time to read this. Upon

a keypress, two options for filling in the blank, *South Station* and *Downtown Crossing*, are displayed below the sentence. They are instructed to choose the best answer and read the complete sentence aloud upon hearing a beep, which is played 1500 milliseconds after *South Station* and *Downtown Crossing* appear. This forces them to make a decision quickly.

We use the term TARGET WORD to refer to the word or phrase that the speaker utters when filling in the blank. We use the term CONTEXT to refer to the fixed part of the utterance. Because the speakers have unlimited time to read over the context before seeing the options for filling in the blank, we consider the target word to be the utterance's LOCUS OF UNCERTAINTY. In this way, we are able to isolate the phrases causing uncertainty (criterion 2).

To elicit both certain and uncertain utterances from each speaker (criterion 1), the transportation questions vary in the amount of real-world knowledge needed to answer the question correctly. Some of the hardest items contain two or three slots to be filled. Because we want the corpus to contain multiple instances of the same word in different contexts (criterion 4), the potential target words are repeated throughout the experiment. This allows us to see whether individual speakers have systematic ways of conveying their level of certainty.

In the vocabulary domain, the speakers are instructed to choose the word that best completes the sentence. To ensure that even the most well-read participants would be uncertain at times (criterion 1), two extra measures were taken. First, the potential target words include three extremely infrequent words: *spoffish*, *hidrotic*, and *vituline*,[1] drawn from a dictionary of obscure words (Byrne, 1974). Second, for five of the twenty items, *none* of the potential target words fit well in the context, generating further speaker uncertainty.

[1] See Appendix A.4 for definitions.

After the speaker reads the sentence aloud, the question, "How certain do you feel about the answer you just gave?" is displayed on the screen. The speaker then indicates their his or her of certainty, with a mouse-click, on a 5-point scale, where 1 is labeled as 'very uncertain' and 5 is labeled as 'very certain.' We refer to this rating as the SELF-REPORTED LEVEL OF CERTAINTY. As we will show in section 2.4, by examining these measures of certainty, our data collection methodology fulfills the crucial criterion (1) of generating a broad range of certainty levels.

The method for eliciting a single utterance is summarized below.

Speech elicitation procedure:

1. The experimenter asks a question *(for transit items only)*.
2. The participant sees the context (i.e., sentence with one or more gaps); the target word options are not shown.
3. The participant sees the target word options below the context. After 1500 ms a beep is played, prompting the participant to read the sentence aloud.
4. The participant rates his or her level of certainty on a 1 to 5 scale.

Participants completed a warm-up activity and two practice items to get acquainted with the procedure before beginning the main body of the experiment. In the body of the experiment, the order of the items was balanced across subjects. To present the materials and record the certainty self-reports, we used the Linger software toolkit (Rohde, 2008).

2.3 Handwritten Digit Domain (Phase 2)

In the second phase of data collection the emphasis was on controlling a speaker's INTERNAL LEVEL OF CERTAINTY. This contrasts the first phase of data collection, where internal certainty depended on the scope of a speaker's vocabulary or his or her knowledge of public transportation in Boston, over which we had no control.

A novel characteristic of the Phase 1 data collection procedure is that we collect level of certainty self-reports during the speech elicitation. We can use these self-reports as a proxy for internal level of certainty if we assume that speakers are honest and accurate in reporting their level of certainty. However, there are two problems with this assumption. First, although the speakers had no incentive to be dishonest, we cannot guarantee that their self-reports are accurate. Second, because internal level of certainty was dependent on a speaker's prior knowledge of public transit routes and the breadth of his or her vocabulary, there was no way, in Phase 1, to verify whether a speaker's self-reported certainty was aligned with his or her actual, internal certainty.

To address these problems, we designed a new set of speech elicitation materials and conducted a second phase of data collection. The design of the Phase 2 materials have the same benefits of the Phase 1 materials, while also allowing us to assess how well level of certainty self-reports reflect a speaker's internal level of certainty.

In this section, we present the Phase 2 data collection method. The method is an adaptation of the Phase 1 method (see section 2.2). The salient difference is that the speech elicitation materials are designed in a way that controls the level of certainty of the stimulus. This is achieved by asking participants to answer questions that necessitate reading handwritten digits that vary in their degree of legibility—some digits are unambiguous and

clear while others are sloppy and hard to decipher.

In section 2.3.1, we discuss our procedure for obtaining the set of handwritten digit images and describe a human computation approach to quantifying each digit's INTRINSIC LEVEL OF CERTAINTY. This allows us to compare a speaker's self-reported certainty to the item's intrinsic level of certainty and verify whether self-reports are a reasonable proxy for internal level of certainty. In section 2.3.2, we present the speech elicitation materials that incorporate these handwritten digit images. We describe the participants in section 2.3.3 and the data collection procedure in section 2.3.4.

2.3.1 Legibility Scores for Handwritten Digits

In this section, we describe how we create speech elicitation materials where each utterance is associated with a fixed, intrinsic level of certainty. We make use of the MNIST database of handwritten digits (LeCun et al., 1998). The MNIST database contains 10,000 handwritten digit images from the United States Postal Service.

The selection of handwritten digits to use in the Phase 2 speech elicitation materials was a stepwise process. This process is summarized below, then each step is described.

Step 1 Use an SVM classifier to identify 400 images (out of all 10,000 images) that may be difficult to read.

Step 2 Generate legibility scores for these 400 images using Mechanical Turk.

Step 3 Select 50 images (out of the set of 400) of varying legibility scores to use in the speech elicitation materials.

In the first step, we use an existing support vector machine classifier (Maji and Malik,

2009) to classify all the images in the MNIST database. This classifier outputs a confidence measure along with the most likely label. We select the 400 images with the lowest confidence measures to use in the subsequent step.

In the second step, we use Amazon's Mechanical Turk (Paolacci et al., 2010; Mason and Suri, 2011) to collect human judgements that we use in generating legibility scores for these 400 images. Mechanical Turk is an online labor market that facilitates the assignment of human workers to quick and discrete *human intelligence tasks* (HITs). We divided the digit images into twenty sections so that each HIT consisted of 20 images. We instructed workers to identify each digit using a drop-down menu. Each digit was labeled by 100 human workers.[2] The full instructions and the parameters of the HIT design are summarized in Appendix A.7. Figure 2.1 shows a screenshot of the Mechanical Turk HIT.

Ensuring worker quality and preventing malicious behavior (e.g., bots written to complete all the HITs in a batch) is a challenge for researchers collecting data on Mechanical Turk (Ipeirotis et al., 2010; Callison-Burch and Dredze, 2010). We took two measures to ensure quality. First, we included a question, such as "What is 4+2?", to verify that the worker was a real person. Second, we randomly included two control images in every HIT. We verified that workers correctly identified those digits before paying them.

We generate a legibility score for each image based on the *entropy* of the human label distribution. In information theory, the entropy of a random variable X is defined as,

$$H(X) = \sum_{i=1}^{N} p(x_i) log p(x_i)$$

[2]The experiment was staged in two rounds, with 10 unique HITs per round. Round 1 took 126 hours (about five days) to complete, with an average time/HIT of 72 seconds. Round 2 took 33 hours (about one and a half days) to complete, with an average time/HIT of 61 seconds. The two experiment rounds were identical in all ways except for the images themselves. We speculate that Round 2 was completed faster than Round 1 due to the time of posting, i.e., weekend vs. weekday.

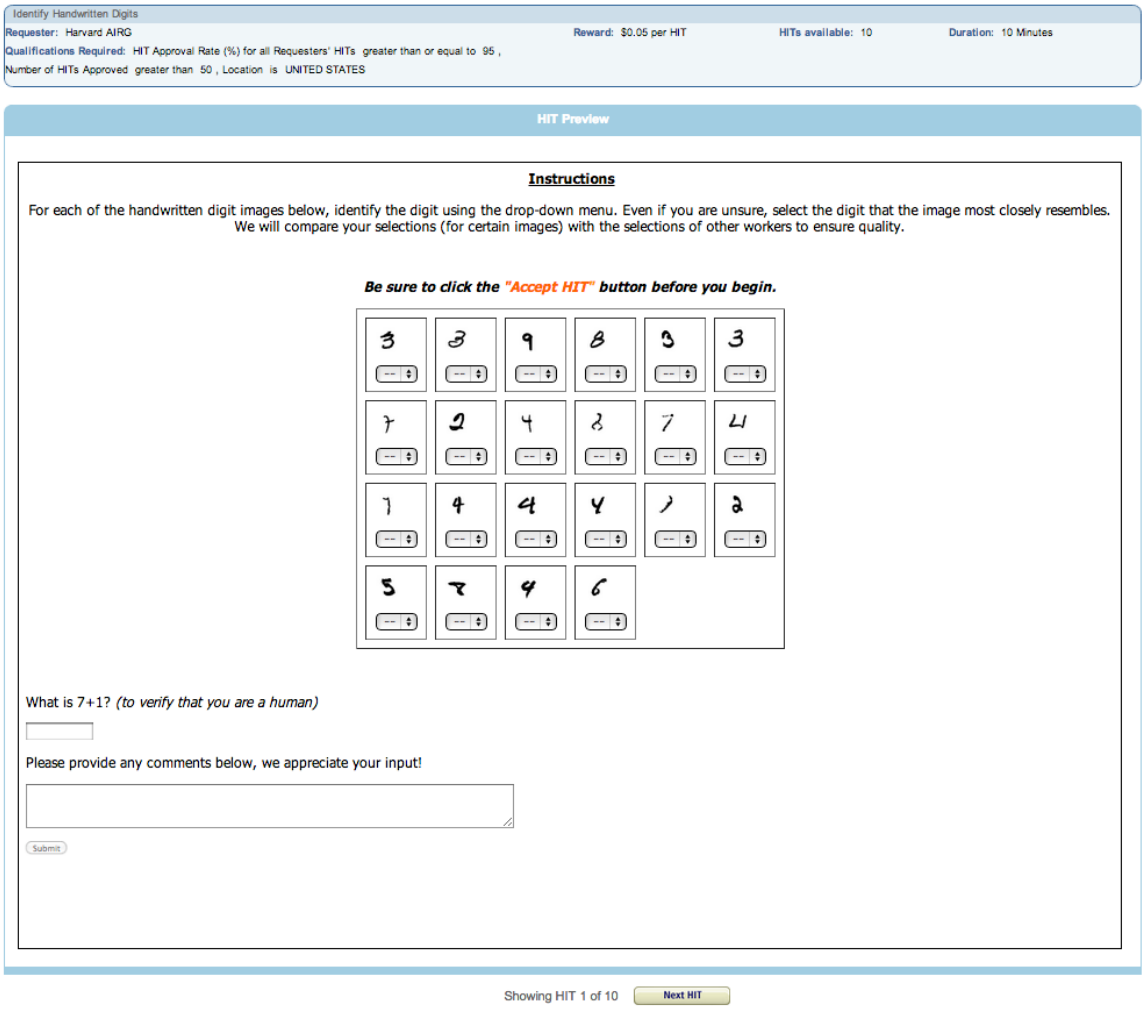

Figure 2.1: Screenshot of the Mechanical Turk HIT for handwritten digit classification.

It is a measure of the uncertainty of a random variable (Cover and Thomas, 1991). For digits that were unambiguous, where all 100 humans agreed on the digit's identity, $H(X) = 0$. For the less legible digits, where workers selected a range of labels, the entropy values were greater than 0. We define the legibility score of an image as $1 - H(X)$. Figure 2.2 shows the frequency of entropy values for the 400 images that were classified by workers on Mechanical Turk.

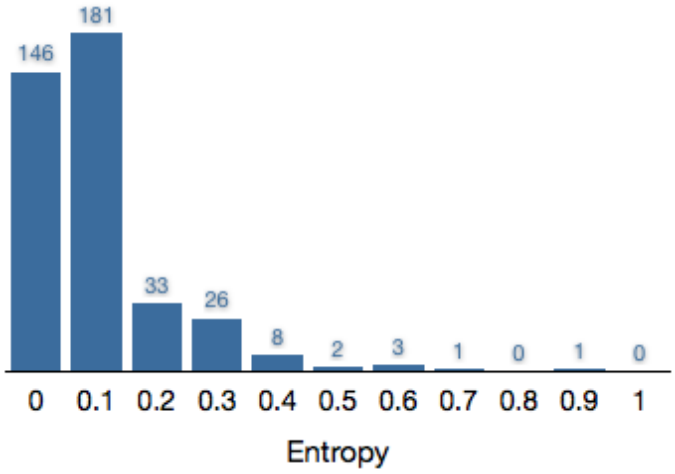

Figure 2.2: The distribution of entropies for the 400 images that were classified by human workers on Mechanical Turk.

While the majority of images were unambiguous or had high agreement, computing the entropies allows us to identify the images that have high ambiguity. Table 2.1 shows three digits of varying legibility, their associated entropies, and the frequencies of the human labels.

In the final step, we select 50 images to use in the speech elicitation stimuli based on the entropies of the human-label distributions. We drew uniformly (as uniformly as possible) from the range of entropies. The resulting set of 50 images is shown in Figure 2.3. The images are displayed from easiest to hardest (low entropy to high entropy) starting from

Table 2.1: Handwritten digit images of varying legibility. We used Amazon's Mechanical Turk to collect 100 human labels for each handwritten digit image. Below each image is the entropy, $H(X)$, of the distribution of human labels, followed by the individual label (x_i) frequencies.

					
$H(X) = 0$		$H(X) = 0.27$		$H(X) = 0.81$	
Label	**Frequency**	**Label**	**Frequency**	**Label**	**Frequency**
'5'	100	'4'	69	'1'	34
		'6'	31	'3'	20
				'5'	15
				'2'	9
				'8'	8
				'7'	5
				'4'	4
				'6'	3
				'0'	2

the top-left and moving left-to-right across the rows. In addition to entropy, we scored the images based on the size of the most commonly labeled class; as one would expect, the two measures were highly negatively correlated ($r = -0.95$).

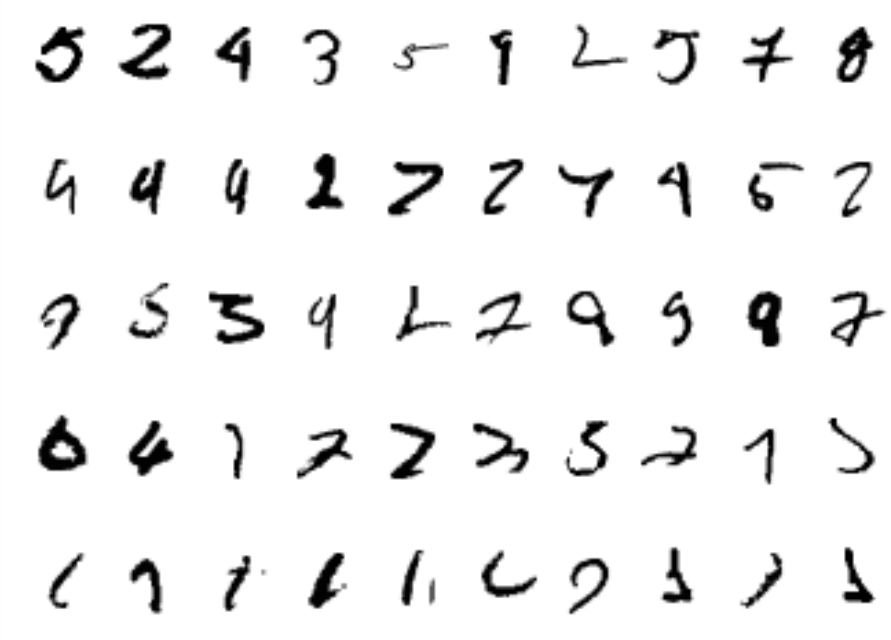

Figure 2.3: Handwritten digit images of varying legibility, ordered from easiest to hardest.

2.3.2 Materials

The materials for eliciting speech were designed so that participants would speak the selected MNIST digit aloud, in the context of answering a question. The handwritten digit images were embedded in an illustration of a train route connecting two U.S. cities. The handwritten digit indicates the train number. Two example train route illustrations are shown in Figures 2.4 and 2.5. The arrow indicates the direction of the train. Beside the departure city and arrival city are clocks indicating the train's departure and arrival times,

respectively.

At the start of the data collection experiment, participants read over a task scenario explaining why they are deciphering handwritten train conductor notes and answering questions about them. The task scenario is included in Appendix A.5. For each train route illustration, participants are asked a single question. There are two forms of questions:

(1) Which train leaves $city_x$ and at what time does it leave?

(2) Which train arrives in $city_x$ and at what time does it arrive?

The participants respond aloud, speaking spontaneously. However, their word choice is influenced by a warm-up task where they are given answers to read aloud. These example answers are of the form:

"Train number 7 leaves Los Angeles at 1:27."

In this way, we are able to control the length and lexical content of the utterances without the participant explicitly reading a sentence aloud.

2.3.3 Participants

Twenty-two members of the Harvard community participated in the data collection. All participants were native English speakers. There were 11 females and 11 males. The mean age was 21.72 (standard deviation 2.97).

2.3.4 Procedure

The procedure for eliciting speech and certainty self-reports is an adaptation of the procedure described in section 2.2.3. As in the first phase of data collection, a beep prompts

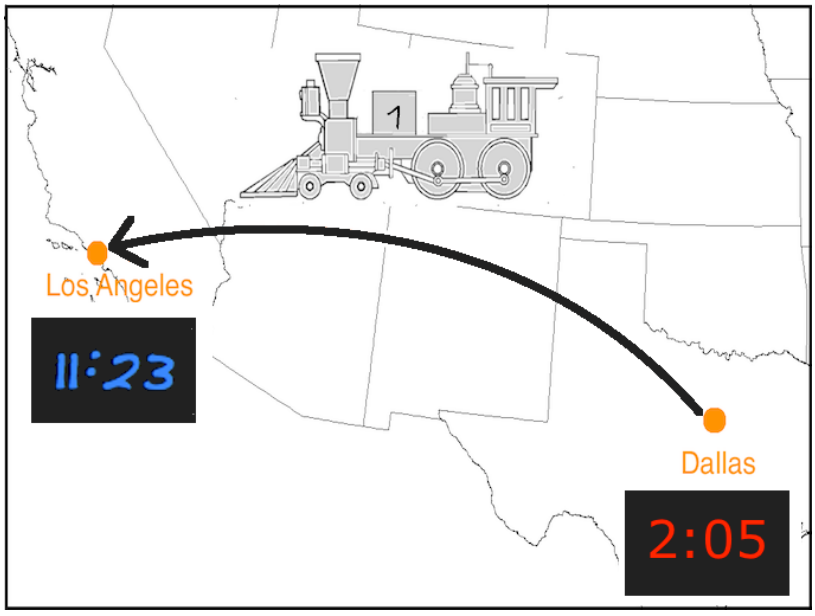

Figure 2.4: Phase 2 speech elicitation materials (a).

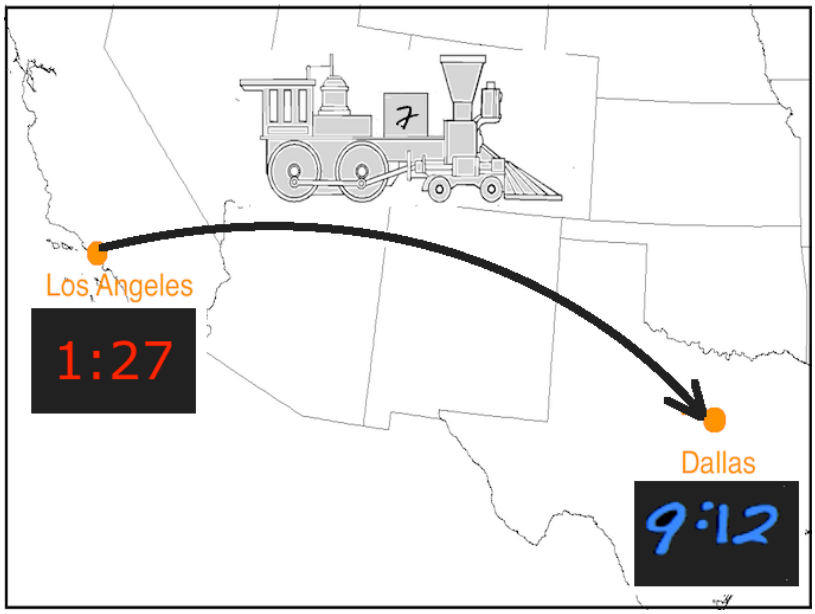

Figure 2.5: Phase 2 speech elicitation materials (b).

the participant to begin speaking. After answering the question, the participant rates his or her level of certainty on a Likert scale (1=very uncertain, 5=very certain). There are two differences: (1) the questions are pre-recorded and integrated into the experiment interface, and (2) the participants answer the questions spontaneously. The procedure is summarized below.

Speech elicitation procedure:

1. The participant is shown a train route illustration.
2. The participant hears a question about the train route.
3. A beep is played, prompting the participant to answer.
4. The participant answers, while viewing the train route illustration.
5. The participant rates his or her level of certainty on a 1 to 5 scale.

Participants completed three warm-up and five practice items to get acquainted with the procedure before beginning the main body of the experiment. In the body of the experiment, the order of the items was random. To present the materials and record the certainty self-reports, we used the PsychoPy software toolkit (Pierce, 2007, 2012).

2.4 Measuring Level of Certainty

The term *level of certainty* has multiple interpretations. Before exploring how to use prosodic information to infer a person's level of certainty, the first step is to make clear how we define level of certainty, and how we measure it.

On the one hand, level of certainty may refer to how certain a person sounds, the *perceived* level of certainty. For the task of inferring level of certainty from prosodic cues, this definition is reasonable — we want our system to hear whatever it is that humans hear. Not surprisingly, this is the definition that has been assumed in previous work on classifying level of certainty (Liscombe et al., 2005). Likewise, the majority of studies on classifying affect also use perceived affect, from the hearer's perspective, as the gold standard (Ang et al., 2002; Litman and Forbes-Riley, 2006).

However, in applications such as spoken tutoring systems (Forbes-Riley et al., 2008) and second language learning systems (Alwan et al., 2007), we would like to know how certain speakers actually are — their *internal* level of certainty, in addition to how certain they are perceived to be. This knowledge affects the inferences such systems can make about the speaker's internal state, for example whether the speaker has a misconception, makes a lucky guess, or might benefit from some encouragement.

Getting a ground truth measurement of a speaker's internal level of certainty is a challenging problem that we address in this thesis. First, we ask speakers to rate their own level of certainty, we refer to this as the *self-reported* level of certainty. We also collect speech data using materials designed to control the speaker's actual, internal level of certainty.

One novel contribution of this thesis is the collection and comparison of these different measures of uncertainty. In the prior work on using prosody to classify level of certainty, no one has attempted to measure a person's internal level of certainty.

Section 2.4.1 presents our approach to measuring certainty from the *hearer's* perspective, including our speech annotation procedure and inter-annotator agreement results. Section 2.4.2 describes our approach to measuring certainty from the *speaker's* perspective.

Observations regarding the comparison of these quantities are presented in section 2.4.3.

2.4.1 Hearer's Perspective

When considering level of certainty from the hearer's perspective, we use the phrase, *perceived level of certainty*. We compute perceived level of certainty by averaging the judgements of a panel of human annotators. The annotators were drawn from the Harvard community. All were native English speakers and had no background in phonetics or linguistics. Five annotators listened to the Phase 1 speech data. Seven annotators listened to the Phase 2 speech data. Every annotator listened to and rated the entire section of the corpus: 600 utterances for Phase 1 and 1100 utterances for Phase 2.

Materials and Procedure

The annotators listened to the audio recordings of the utterances in a random order in twelve sections. We instructed the annotators to rate *how certain the speaker sounded* regardless of how sensible the resulting sentence was. The annotators did not see any contextual information such as the instructions given to the speakers, the questions, the possible target words, or the locations of the target words. The annotators rated level of certainty for each utterance on the same 5-point scale used for the self-reports (1 = very uncertain, 5 = very certain). The instructions for the annotators are included in Appendix A.2 (Phase 1) and Appendix A.6 (Phase 2).

Inter-annotator Agreement

Inter-annotator agreement was calculated using Cohen's Kappa statistic (κ), a common measure of categorical agreement. The Kappa statistic is defined as,

$$\kappa = \frac{Pr(\text{observed agreement}) - Pr(\text{agreement by chance})}{1 - Pr(\text{agreement by chance})}$$

Kappa values can range from $[-1, 1]$, where $\kappa = 1$ is perfect agreement, $\kappa = -1$ is perfect disagreement, and agreement by chance results in $\kappa = 0$. Because perception of affect and emotion is very subjective, we do not expect to obtain perfect agreement. Related studies in the area of emotion detection report Kappa scores from $0.4 - 0.6$ (Liscombe et al., 2005; Litman and Forbes-Riley, 2006).

We compute agreement between each pair of annotators then average the scores. We compute Kappa scores for two cases. In the first case, we consider five classes (the five possible certainty ratings). In the second case, we convert these five classes to three classes (uncertain, neutral, certain), to compare our agreement levels with prior work (Liscombe et al., 2005). In this latter case, we map the 5-classes to 3-classes in this way: uncertain = 1 or 2, neutral = 3, certain = 4 or 5. This partition maximizes inter-annotator agreement.

For Phase 1, in the case where we consider five classes, the average Kappa score between annotators is 0.284. In the case where we consider three classes, the average Kappa score between annotators is 0.45. This Kappa of 0.45 is on par with the 0.52 average Kappa score reported by Liscombe et al. (2005).

For Phase 2, in the case where we consider five classes, the average Kappa score between annotators is 0.129. In the case where we consider three classes, the average Kappa score between annotators is 0.223.

Perceived Levels

Figure 2.6 shows the distribution of self-reports for each of the three domains: vocabulary, transportation, and handwritten digits. We see that in all three domains the histograms have a similar shape, with a mode of 4.

2.4.2 Speaker's Perspective

From the speaker's perspective, there are two categories of certainty measurements that we compute. First, speakers rate their own level of certainty, as described in section 2.2. Second, for the utterances in the handwritten digit domain (Phase 2), we estimate internal level of certainty based on the digit image legibility scores derived from the Mechanical Turk human judgement data (as described in section 2.3.1).

Self-reported Certainty

Figure 2.7 shows the distribution of self-reports for each of the three domains: vocabulary, transportation, and handwritten digits. In the vocabulary domain, there is a distinct negative slope; speakers felt uncertain more often than not. In the transportation domain, speakers were certain more often than they were uncertain, though in either a high or low extreme more often than neutral. In the digits domain, there is a distinct positive slope; speakers felt certain most of the time.

Internal Certainty

Although measuring a speaker's actual internal level of certainty is impossible, the Phase 2 data collection materials were designed in an attempt to control internal level of

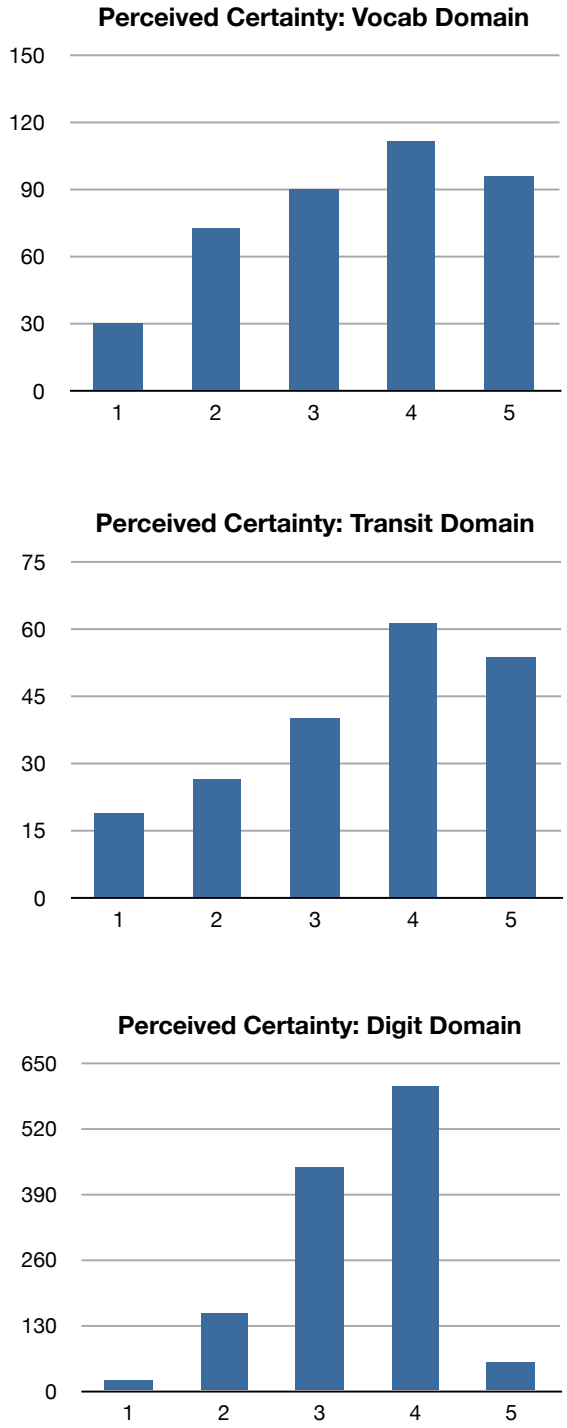

Figure 2.6: The distributions of perceived certainty ratings in the Uncertainty Corpus for each of the three domains: Vocabulary (top), Transportation (middle), and Digits (bottom). 1=very uncertain, 5=very certain.

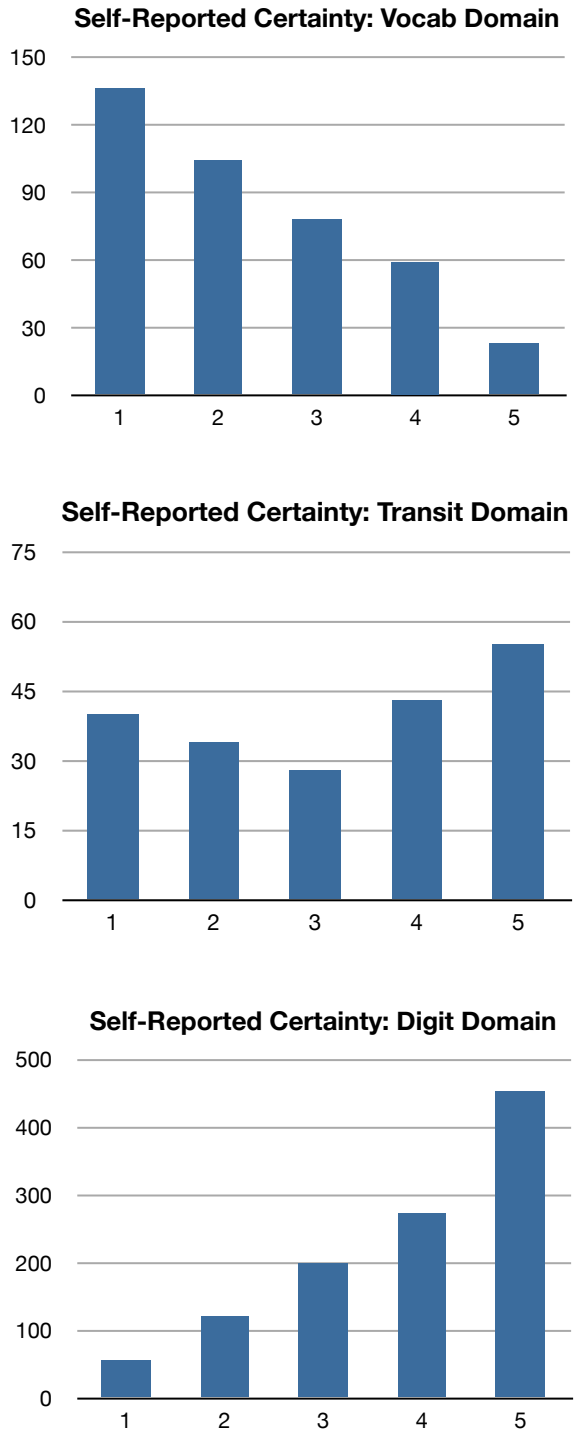

Figure 2.7: The distributions of self-reported certainty ratings in the Uncertainty Corpus for each of the three domains: Vocabulary (top), Transportation (middle), and Digits (bottom). 1=very uncertain, 5=very certain.

certainty. The Phase 2 materials incorporate images of handwritten digits (see section 2.3). We assign each image a legibility score, based on annotations from 100 humans, then use this score as a measure of the image's *intrinsic level of certainty*. Figure 2.8 shows the distribution of legibility scores for the fifty digit images in the Phase 2 data collection materials.

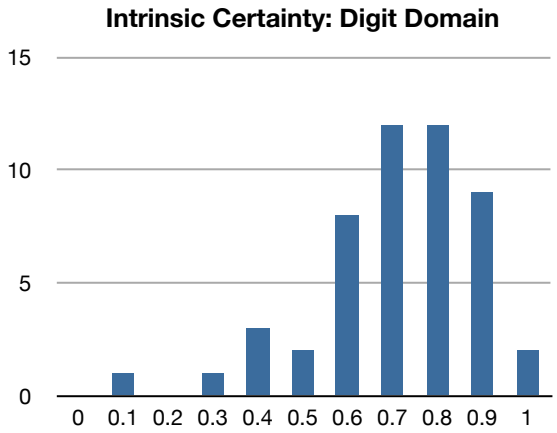

Figure 2.8: The distribution of legibility scores for the fifty digit images in the Phase 2 data collection materials.

2.4.3 Observations

A key contribution of this work is our attention to internal level of certainty. Because we elicit self-reported certainty ratings for all of the utterances in our corpus (see section 2.4.2), we are able to make observations regarding two aspects of cognitive state: *transparency* and *self-awareness* (as introduced in section 1.2). In this section, we discuss our observations of speaker transparency and speaker self-awareness with respect to the first phase of the data collection (vocabulary and transportation domains). These observations are discussed in previously published work (Pon-Barry and Shieber, 2010).

Transparency

Because our corpus contains both external (perceived by a panel of human judges) and internal (self-reported by the speaker) measures of certainty, we are able to assess a speaker's transparency. For a single utterance, a speaker is transparent if the perceived level of certainty is an accurate reflection of his or her self-reported level of certainty. The concept of transparency was introduced in section 1.2.

In the Uncertainty Corpus (Phase 1), speakers are transparent for 64% of the utterances. The correlation coefficient between perceived and self-reported certainty is 0.42. We saw in Figure 2.6 that across all domains, the distribution of perceived certainty ratings were concentrated on the certain side. On the other hand, Figure 2.7 showed that the distributions of self-reported certainty ratings took different shapes depending on the domain. We can see the *relative* frequencies of these external and internal certainty ratings in the heat map in Figure 2.9. Figure 2.9 shows the relative frequencies for the Phase 1 data. The concentration of darker squares above the diagonal line correspond to the cases where the annotators (human listeners) perceived the speakers as being more certain than the speakers themselves self-reported.

Of the 600 utterances in the Phase 1 portion of the corpus, 41% have perceived certainty levels that are more than one unit *greater* than the self-reported level.[3] On the other hand, only 8% of the utterances have perceived certainty levels more than one unit *less* than the self-reported rating.

[3]We considered the possibility that there is a gender effect related to our observation that perceived certainty ratings are consistently greater than self-reported certainty ratings. While our preliminary analysis suggests this may be the case, the Phase 1 participants were not balanced for gender. The sample size, 6 males and 14 females, limited our ability to draw statistically significant conclusions. The participants in Phase 2 of the data collection were balanced for gender; in our future work, we plan to investigate whether there are statistically significant gender-related effects.

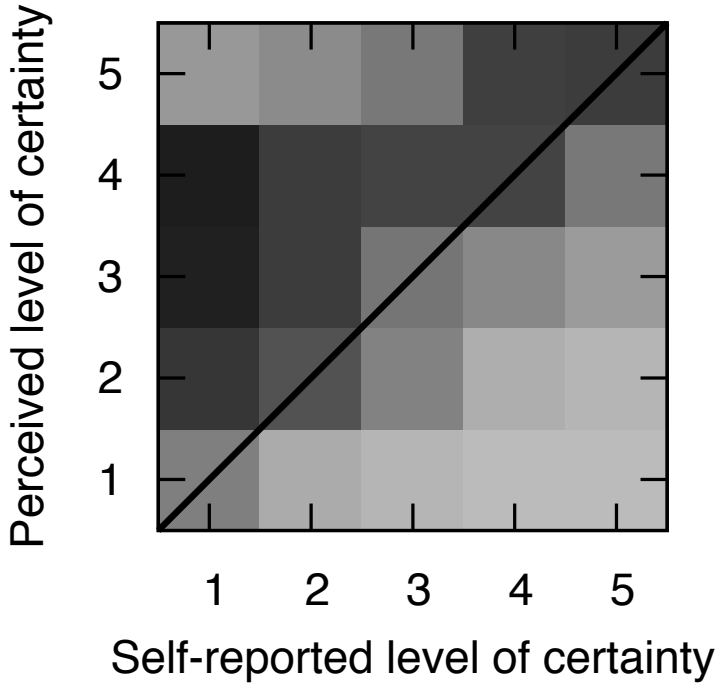

Figure 2.9: Heat map illustrating the relative frequencies of Phase 1 utterances grouped according to self-reported level of certainty and (quantized) perceived level of certainty. Darker squares indicate greater frequency.

Looking at average transparency per individual speaker can shed light on whether speaking personalities are present among the speakers in our corpus. The data shows a wide range of variation among individual speakers. Average transparency values for individuals range from 27% to 80%, with a median of 67%. Some individual speakers were consistently transparent, while others had mixed degrees of transparency. In other cases, individuals sounded very certain all of the time—even when they felt uncertain. For example, a few individual speakers were consistently perceived as uncertain, regardless of their self-reported internal certainty. There are many factors that can affect how transparent a person is, for example, related work in psychology argues that people's beliefs about their transparency and thus the emotions they convey are highly dependent on the context of the interaction (Parkinson, 2008).

When we go one step further and examine average individual transparency for correct versus incorrect utterances, we see that some speakers are more transparent when correct than when incorrect. This is illustrated in Figure 2.10, where each dot corresponds to an individual speaker. If all individuals were consistently transparent (or not transparent) regardless of their correctness, we would expect them to fall along the dashed line. The presence of outliers and clusters of individuals in Figure 2.10 suggests that there may be different 'speaking personalities' related to transparency, similar in spirit to the personality types described by Mairesse et al. (2007). If such speaking personality clusters exist, identifying them may be useful for dialogue systems or other applications that wish to accurately model a user's internal level of certainty.

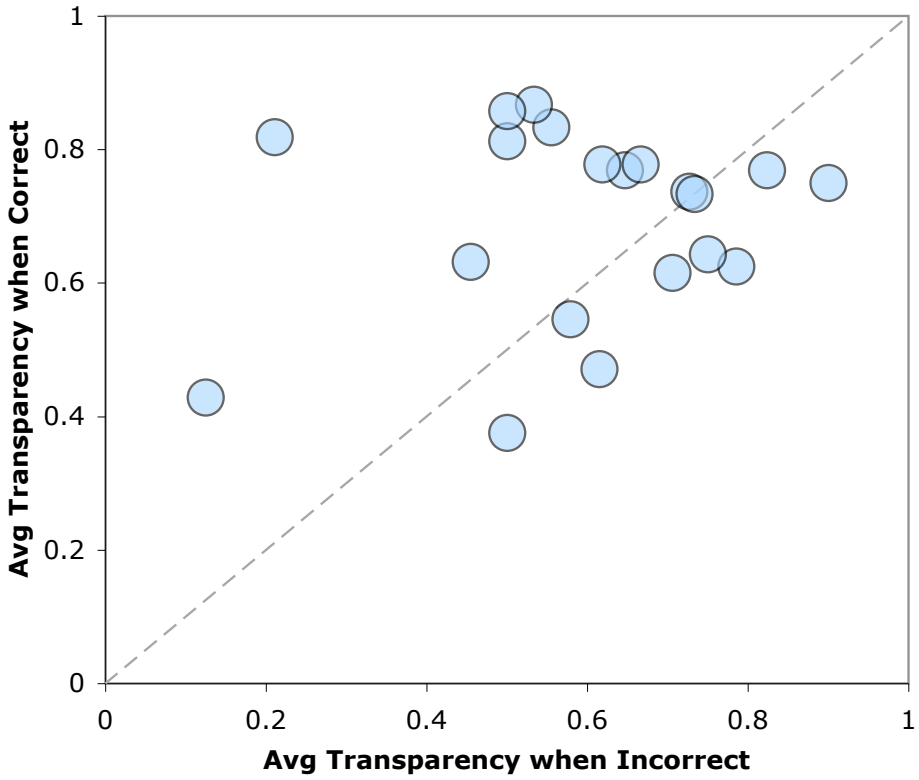

Figure 2.10: Average transparency when incorrect versus correct. Each dot represents a single person.

Self-awareness

Because our corpus contains certainty self-reports and the utterances in the corpus can be labeled as correct or incorrect answers, we are able to assess a speaker's self-awareness. The concept of self-awareness was introduced in section 1.2. For a single utterance, we consider a speaker to be self-aware if:

- his or her answer is correct AND self-reported certainty ≥ 3 *or*
- his or her answer is incorrect AND self-reported certainty ≤ 3

Over the whole Uncertainty Corpus (Phase 1), speakers were self-aware for 73% of the responses. Looking at individual speakers, the average self-awareness percentages range from 43% to 87%, with a median of 72%.

Figure 2.11 shows the distribution of self-reported certainty for correct answers; Figure 2.12 shows the distribution of self-reported certainty for incorrect answers. The general upward slope in Figure 2.11 and downward slope in Figure 2.12 are in accordance with our expectations, given that the speakers were self-aware for 73% of their responses. The cases that we are most interested in are the incongruous edge cases — when the speaker is correct and reports feeling *uncertain* (a level of 1 or 2) or when the speaker is incorrect and reports feeling *certain* (a level of 4 or 5). These are the cases where an adaptive dialogue system would alter its default dialogue strategy, as discussed in section 1.2. Looking at Figures 2.11 and 2.12, we see that there were more instances of the former category, being correct and feeling uncertain, than the latter category in the Uncertainty Corpus.

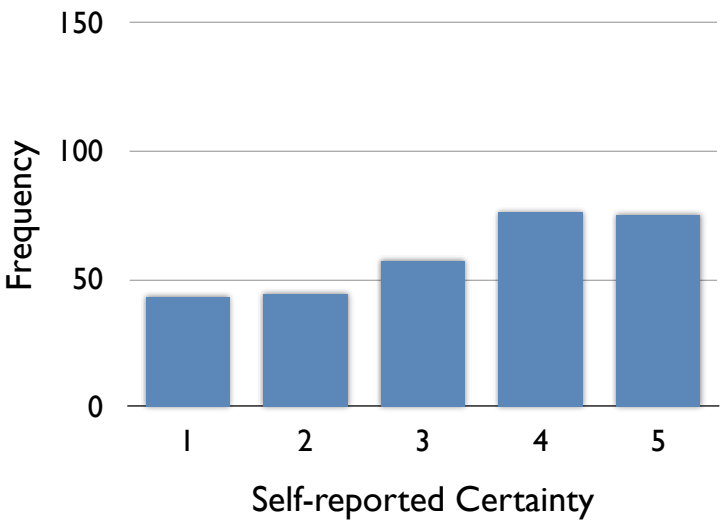

Figure 2.11: Self-reported certainty ratings for *correct* answers.

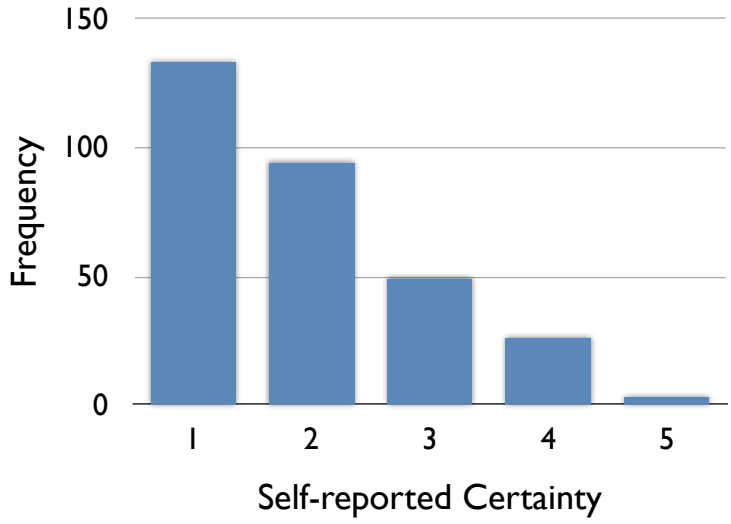

Figure 2.12: Self-reported certainty ratings for *incorrect* answers.

2.5 Corpus Statistics

The Uncertainty Corpus contains speech recordings, level of certainty annotations, and acoustic feature vector data. The speech elicitation materials include items from three domains: vocabulary, public transportation, and handwritten digits. The complete materials are in Appendix A.

In total, the Uncertainty Corpus has 1876 utterances and 148.79 minutes of speech. Table 2.2 lists these statistics for each phase of the data collection.

The speech recordings are available upon request, for research purposes only. The level of certainty annotations and acoustic feature vector data are available for download through the Dataverse Network.[4]

Table 2.2: Descriptive statistics for the Uncertainty Corpus.

	Number of Utterances	**Minutes of Speech**
Phase 1	600	59.48
Phase 2	1100	89.31
Total	1876	148.79

2.6 Chapter Summary

The design of our data collection method achieves several goals. First, we control the difficulty of the items, thus ensuring varying levels of certainty. Second, we isolate the

[4]http://dvn.iq.harvard.edu/dvn/dv/ponbarry

locus of uncertainty. Third, we achieve lexical consistency across utterances in the corpus, minimizing the non-affective factors that influence prosody. Fourth, we collect multiple instances of a small set of target words, allowing for word-level prosodic analysis. Fifth, we elicit self-reported measures of internal certainty directly from the speakers (in addition to eliciting judgements from a panel of human listeners). These characteristics distinguish this work from of the prior work on inferring speaker affect.

In section 2.3, we presented a set of data collection materials that allow us to attempt to control a speaker's internal level of certainty.

Our corpus demonstrates that there are systematic differences between perceived certainty and self-reported certainty. Research that treats them as equivalent quantities may be overlooking significant issues. Further, we discussed how knowing a user's internal and perceived levels of certainty allows us to assess higher-order mental states such as self-awareness and personality characteristics such as transparency. When planning an adaptive dialogue, the default strategy is to assume that users are both self-aware and transparent. However, if an intelligent system has evidence suggesting the user is *not* self-aware or is *not* transparent, it can use this knowledge to remediate misconceptions, offer encouragement, or determine the right feedback to give.

Chapter 3

Prosodic Analysis

In this chapter, we give an overview of our approach to selecting the features that we use to represent the prosody of an utterance. We begin, in section 3.1, by summarizing theoretical work in the area of intonational meaning, in other words, theories of how intonation contours are used in natural language conversation. These theories provide motivation for our task of inferring affect based on prosody. In section 3.2, we describe a set of low-level acoustic features that we use to represent the prosody of an utterance. We require these features to be easily computed from the speech signal because our next step, in chapter 4, will be to perform machine learning classification experiments. The features presented in section 3.2 fall into three groups: pitch features, loudness features, and temporal features. In section 3.3, we discuss the relationship between these features and level of certainty in the Uncertainty Corpus. Lastly, we compare our prosodic feature selection to related work in section 3.4.

3.1 Theories of Prosody

Although early characterizations of prosody suggested that prosodic structure was inferable from the syntactic structure of an utterance, it has since been argued that the prosodic structure of an utterance is distinct from the syntactic structure: many aspects of spoken utterances cannot be predicted by (morpho-)syntactic structure, and syntax does not necessarily constrain prosodic structure (Shattuck-Hufnagel and Turk, 1996).

Therefore, two central questions in the study of prosody are: (1) how do we represent prosodic structure? And (2) what meaning is associated with particular prosodic patterns? In other words, why do speakers produce different intonational contours and how do listeners interpret them? In the descriptions below, we refer to pitch accents that have a rise or a fall. This refers to the slope of the pitch over a small time window—whether the fundamental frequency of the speaker's voice is increasing or decreasing.

Bolinger (1989) was one of the first linguists to provide a characterization of intonational contours and pitch accents. He defines six intonational profiles: three basic profiles, and three compositional profiles.

- Profile *A*: abrupt fall down *from* accented syllable
- Profile *B*: rise up to accented syllable
- Profile *C*: abrupt fall down *to* accented syllable
- Profile *CA*: a low-high-low profile, with the accented syllable at the start or throughout
- Profile *AC*: an abrupt fall down from an accented syllable with a rise after the fall

- Profile *CB*: a "low-high-slither" profile, with the accented syllable at the start ("slither" describes a mainly flat but slightly downward slope)

Bolinger makes observations about the meaning of each profile. For example, he proposes that a profile *A* often signals new information. Bolinger uses a unique style of descriptive notation in which the vertical typesetting of words reflects the intonational contours. An example of profile *A*, an abrupt fall from an accented syllable, is shown in Figure 3.1. An example of profile *AC*, an abrupt fall from an accented syllable with a rise after the fall, is shown in Figure 3.1.

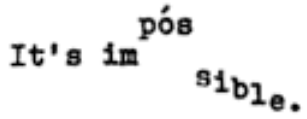

Figure 3.1: Intonational profile *A* of Bolinger (1989): an abrupt fall from an accented syllable.

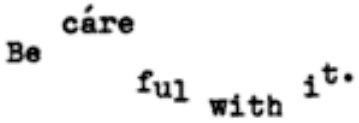

Figure 3.2: Intonation profile *AC* of Bolinger (1989): an abrupt fall from an accented syllable with a rise after the fall.

Although Bolinger's style of presentation was not widely adopted, his idea of systematizing intonational contours shares similarities with the compositional theory of "tune" interpretation of Pierrehumbert and Hirschberg (1990), which is the basis for the widely-used ToBI (Tones and Break Indices) descriptive system (Silverman et al., 1992).

Pierrehumbert and Hirschberg's theory of "tune" interpretation is grounded in Grosz and Sidner's model of discourse processing (Grosz and Sidner, 1986). They propose that

different intonational contours reflect changes to the *intentional state* and the *attentional state* of the discourse. In this model of discourse, the speaker and the hearer each have a set of beliefs. Mutual beliefs are the intersection of these beliefs, from the perspective of either the speaker or the hearer (as neither has full knowledge of the other's beliefs). Pierrehumbert and Hirschberg propose that the meaning of a particular pitch accent can be expressed in terms of the speaker's goal of adding something to the mutual beliefs of the listener. A central part of Pierrehumbert and Hirschberg's proposal is the idea that tune meaning is compositional. In their proposal, there are three types of tones: pitch accents, phrase accents, and boundary tones.

Pierrehumbert and Hirschberg (1990) assume that there are six different types of pitch accents in English. These six types and their proposed meanings are summarized below (L = low tone, H = high tone, * = stressed syllable, and % = boundary tone). A significant difference between this work and previous work (e.g., Bolinger (1989), is that the basic building blocks are in terms of pitch height (L and H), rather than in terms of slope ('rise' and 'fall'). The central points of their theory are outlined below.

General Rule: All pitch accents make the accented material salient.

H*: items made salient by H* are to be treated as *new* in the discourse

L*: items made salient by L* are items that the speaker intends to be salient but not to be part of what the utterance is predicating.

L+H: these are used by speakers to convey the salience of some scale (i.e., partial ordering) linking the accented item to other salient items in the (hearer's) mutual beliefs.

L*+H: evokes a scale, and conveys lack of predication.

L+H*: evokes a scale, and conveys that the accented item (not some other salient item) should be mutually believed

H+L: these indicate that the hearer should be able to infer support for an instantiation, based on the hearer's representation of the mutual beliefs.

H*+L: evokes support for an instantiation, and makes a predication (Note: Pierrehumbert and Hirschberg note that H*+L accents are very rare).

H+L*: evokes support for an instantiation, but does not makes a predication.

The proposals of Bolinger and of Pierrehumbert and Hirschberg address the topic of intonational contours. Another aspect of prosody is the placement of intonational boundaries. While intonational boundaries do not signal a variety of meanings as intonational contours do, they define the prosodic constituents of an utterance, which directly influences the interpretation of the utterance.

Pierrehumbert and Hirschberg's theory evolved into the ToBI (Tones and Break Indices) labeling system (Silverman et al., 1992), a framework for describing prosodic events at the perceptual level. For many years, getting ToBI labels required significant manual annotation by skilled human annotators. Recent advances have been made in automatically producing ToBI labels from the speech signal (Ananthakrishnan and Narayanan, 2008; Rosenberg, 2009), but the state of the art is still below human-level: these systems do well on prosodic phrase boundary detection but not as well on pitch accent classification.

On the other hand, several studies have focused on low-level acoustic features corresponding to pitch, loudness, and timing, since these features are easy to extract from the speech signal (Ang et al., 2002; Fernandez, 2004; Liscombe et al., 2005; Litman and

Forbes-Riley, 2006; Iseli et al., 2006). We follow this approach here.

3.2 Computation of Prosodic Features

The utterances in the Uncertainty Corpus were segmented manually. We consider the start of the utterance to be when the participant was prompted to begin speaking. The participants were instructed to start speaking immediately after hearing a beep (see section 2.2.3). If participants paused before beginning to speak, we include these pauses as part of the utterance. We consider the end of the utterance to be when the participant stopped speaking.

The low-level prosodic features that we use in our experiments fall into three groups: pitch features, loudness features, and temporal features. This set of features was selected in order to be comparable with the set of prosodic features used in the experiments of previous work on certainty classification (Liscombe et al., 2005). Unlike this previous work, we compute contextual prosodic features—features relative to the *target word* and *context* segments, in addition to features at the utterance-level.

When creating the target word segments, utterance-internal pauses were grouped with the word they preceded. That is, a pause immediately preceding the target word would be grouped with the target word; a pause immediately following the target word would be grouped with the context.

3.2.1 Pitch Features

When we talk about the pitch in spoken language, we are talking about the FUNDAMENTAL FREQUENCY (F0) of the waveform. Speech waveforms are complex—they are composed of multiple periodic waves. The fundamental frequency is the rate at which the largest periodic wave repeats. Pitch estimation is an active area of research.

Below are the pitch features that we compute. The pitch features are represented as z-scores ($\mu = 0$, $\sigma^2 = 1$) normalized by speaker. We use the Wavesurfer toolkit (Sjölander and Beskow, 2000, 2008), with a window size of 200 ms, to estimate the F0 values. We use the the Praat toolkit (Boersma and Weenink, 2010) to compute the absolute slope feature.

- **Min F0**: minimum F0 value in segment
- **Max F0**: maximum F0 value in segment
- **Mean F0**: mean F0 value in segment
- **Stdev F0**: standard deviation of the F0 values in segment
- **Range F0**: max F0 $-$ min F0
- **Relative position min F0**: time point that min F0 occurs / length of segment
- **Relative position max F0**: time point that max F0 occurs / length of segment
- **Absolute slope (Hz)**: a measure for the average local variability in F0 (Hz) for all time points t_i:

$$\frac{1}{N}\sum_{i}^{N}\frac{|F0(t_i) - F0(t_{i-1})|}{t_i - t_{i-1}}$$

Consider two examples from the Uncertainty Corpus, shown in Figures 3.3 and 3.4. These two figures correspond to two different speakers in our corpus uttering, *"She's a redoubtable opponent."* They were generated with the Praat toolkit. The three horizontal sections in each figure from top to bottom are: the waveform, the spectrogram, and the words. The blue line segments in the middle section represent the estimated pitch (fundamental frequency). In Figure 3.3, the speaker has a *high pitch accent* on the word "redoubtable" whereas in Figure 3.4, the word "redoubtable" is unaccented.

3.2.2 Loudness Features

We represent loudness by computing the RMS (root mean squared) amplitude of the waveform. RMS normalization is common in practice; it gives a measure of loudness while minimizing the effects of any differences in setup, such as the distance between the speaker and the microphone. The feature values are computed with the Wavesurfer toolkit. The features are represented as z-scores ($\mu = 0$, $\sigma^2 = 1$) normalized by speaker.

- **Min RMS**: minimum RMS value in segment
- **Max RMS**: maximum RMS value in segment
- **Mean RMS**: mean RMS value in segment
- **Stdev RMS**: standard deviation of the RMS values in segment
- **Relative position min RMS**: time point that min RMS occurs / length of segment
- **Relative position max RMS**: time point that max RMS occurs / length of segment

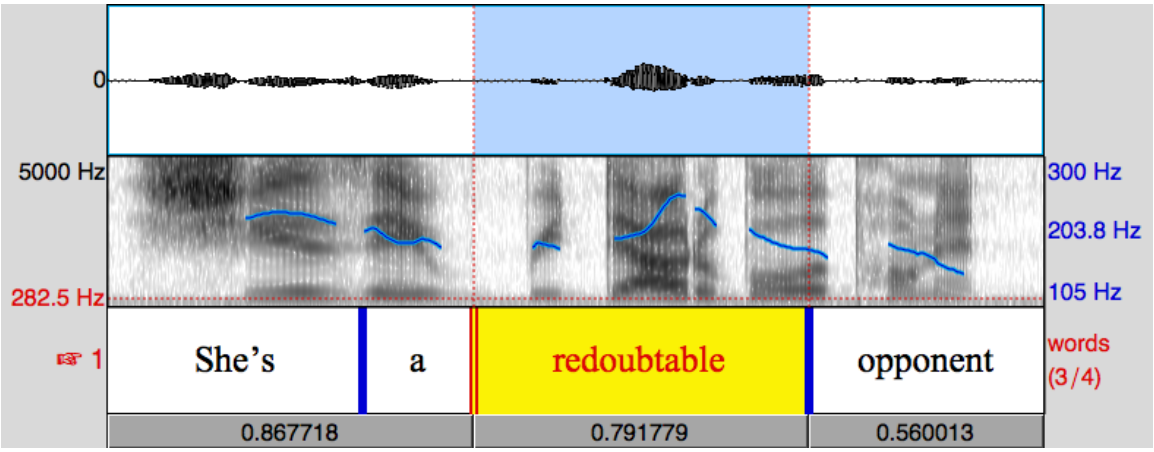

Figure 3.3: A high pitch accent on the word "redoubtable". The blue line segments represent the estimated pitch (fundamental frequency).

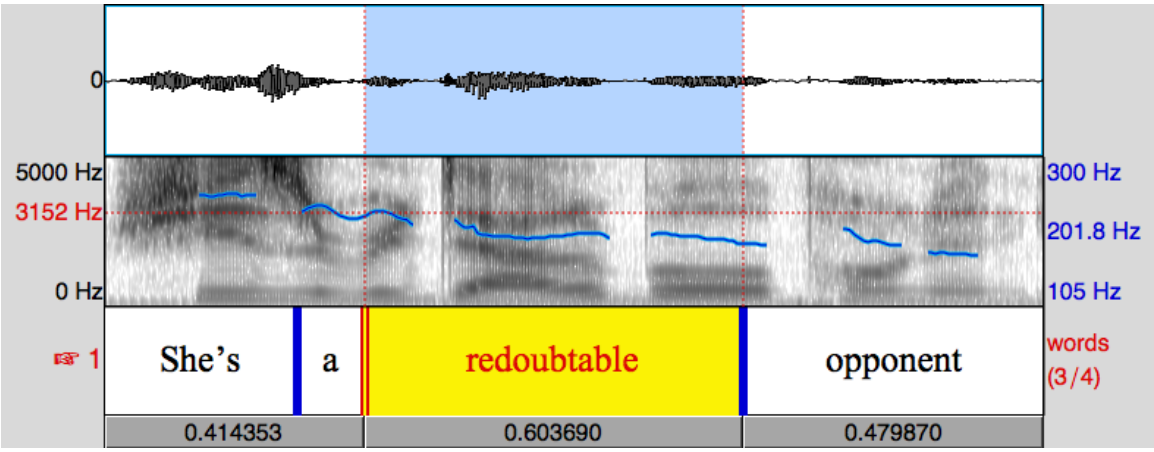

Figure 3.4: A non-accented example of the word "redoubtable". The blue line segments represent the estimated pitch (fundamental frequency).

3.2.3 Temporal Features

We compute five temporal features, listed below. Unlike the pitch and intensity features, the temporal features are not normalized by speaker. We define a pause to be any stretch of silence lasting more than 500 milliseconds. The speaking rate feature is defined as the number of syllables in the segment divided by the speaking duration.

- **Total duration**: total duration of the segment
- **Total silence**: cumulative duration of pauses in the segment
- **Percent silence**: total silence / total duration
- **Speaking duration**: total duration − total silence
- **Speaking rate**: number of syllables / speaking duration

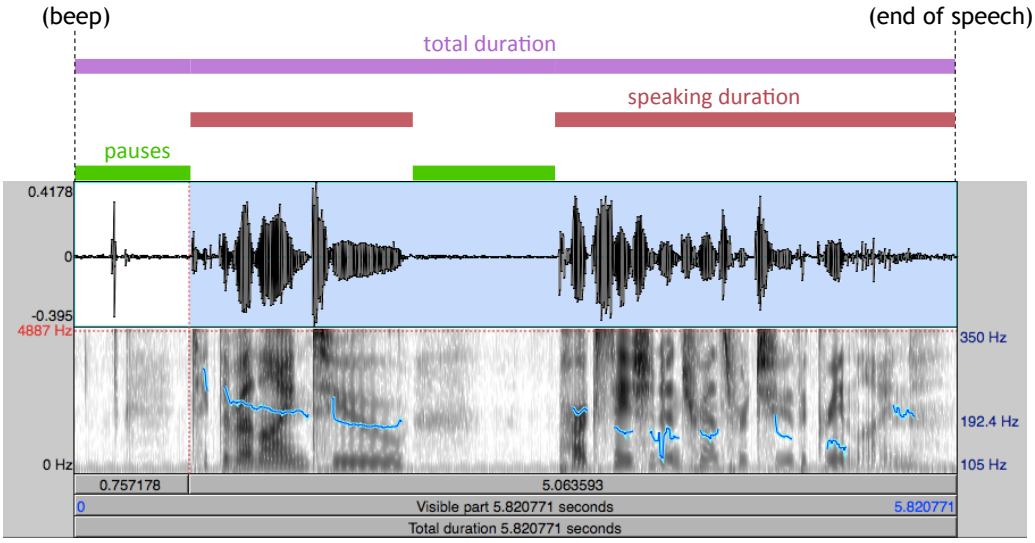

Figure 3.5: Illustration of temporal features.

Figure 3.5 shows an example utterance, with horizontal bars indicating the duration features. The corresponding sentence is, *"Take the Red Line to... Park Station and transfer to the Green Line."*

3.3 Prosody and Level of Certainty

We discuss observations regarding the prosodic features described in section 3.2 with respect to perceived and self-reported levels of certainty in section 3.3.1 and section 3.3.2, respectively.

3.3.1 Relationship with Perceived Certainty

Table 3.1 shows the correlations between perceived certainty and prosodic features for whole utterances, target word segments, and context segments. Statistical significance is indicated with asterisks: * $p < 0.05$, ** $p < 0.01$. The correlations between average rating and prosodic features extracted from whole utterances suggest that temporal features (i.e., total silence, percent silence, total duration, speaking duration) are the features most strongly associated with the perceived level of certainty. Other features, including absolute slope F0, range F0, and speaking rate, had statistically significant but smaller correlations with the perceived certainty.

We observed that some features, such as absolute slope F0, have stronger correlations in the whole utterance than in the context or target word. For features behaving in this way (absolute slope F0, total silence, total duration, speaking duration, speaking rate), separating the context from the target word does not provide any additional information.

Table 3.1: Correlations between *perceived* certainty and prosodic features for whole utterances, context segments, and target word segments.

Feature	Whole Utterance	Context	Target Word
Min F0	0.107^{*}	0.119^{**}	0.041
Max F0	-0.073	-0.153^{**}	-0.045
Mean F0	0.033	0.070	-0.004
Stdev F0	-0.035	-0.047	-0.043
Range F0	-0.128^{**}	-0.211^{**}	-0.075
Rel. position min F0	0.042	0.022	0.046
Rel. position max F0	0.015	0.008	0.001
Absolute slope F0	0.275^{**}	0.180^{**}	0.191^{**}
Min RMS	0.101^{*}	0.172^{**}	0.027
Max RMS	-0.091^{*}	-0.110^{*}	-0.034
Mean RMS	-0.012	0.039	-0.031
Stdev RMS	-0.002	-0.003	-0.019
Rel. position min RMS	0.101^{*}	0.172^{**}	0.027
Rel. position max RMS	-0.039	-0.028	-0.007
Total silence	-0.643^{**}	-0.507^{**}	-0.495^{**}
Percent silence	-0.455^{**}	-0.225^{**}	-0.532^{**}
Total duration	-0.592^{**}	-0.502^{**}	-0.590^{**}
Speaking duration	-0.430^{**}	-0.390^{**}	-0.386^{**}
Speaking rate	0.090^{*}	0.014	0.136^{**}

More interestingly, we observed that features such as range F0 have stronger correlations in the context than in the whole utterance or the target word. This suggests that as a cue to uncertainty, the range F0 feature is manifested most strongly in the context.

We observe that the percent silence feature has a much stronger correlation in the target word than in the context. This suggests that, as a cue to uncertainty, the percent silence feature is manifested most strongly in the target region. The opposite holds for features such as speaking duration, range F0, and min RMS. That is, these features have much stronger correlations in the context region than in the target word region.

3.3.2 Relationship with Self-Reported Certainty

Table 3.2 shows the correlations between self-reported certainty and prosodic features for whole utterances, target word segments, and context segments. Statistical significance is indicated with asterisks: * $p < 0.05$, ** $p < 0.01$.

3.4 Related Work

The decision of which acoustic features to include in our experiments is motivated by the goal of using features comparable to existing related work. Our set of pitch, intensity, and temporal features is identical to those that were used by Liscombe et al. (2005) in their work on classifying level of certainty in tutorial dialogues. Liscombe et al. (2005) also include dialogue turn-related features in their classification experiments.

Prior studies on classifying positive and negative emotion in speech use a similar set of prosodic features (Ang et al., 2002; Lee and Narayanan, 2005). Lee and Narayanan

Table 3.2: Correlations between *self-reported* certainty and prosodic features for whole utterances, context segments, and target word segments.

Feature	Whole Utterance	Context	Target Word
Min F0	0.079	0.289^{**}	-0.225^{**}
Max F0	-0.032	-0.116^{**}	-0.109
Mean F0	-0.012	0.185^{**}	-0.193^{**}
Stdev F0	0.026	-0.113	0.057
Range F0	-0.066	-0.250^{**}	0.087
Rel. position min F0	0.078	0.039	0.084
Rel. position max F0	0.046	-0.067	0.115^{*}
Absolute slope F0	0.170^{**}	0.111	0.029
Min RMS	0.178^{**}	0.341^{**}	-0.085
Max RMS	-0.189^{**}	0.166^{**}	-0.107
Mean RMS	0.038	0.224^{**}	-0.190^{**}
Stdev RMS	-0.103	-0.060	-0.058
Rel. position min RMS	0.061	-0.058	0.086
Rel. position max RMS	0.015	0.009	-0.025
Total silence	-0.400^{**}	-0.370^{**}	-0.190^{**}
Percent silence	-0.306^{**}	-0.139^{*}	-0.233^{**}
Total duration	-0.506^{**}	-0.513^{**}	-0.112
Speaking duration	-0.460^{**}	-0.474^{**}	0.073
Speaking rate	0.063	0.105	-0.091

(2005) also include formant-related features (a formant is a resonant frequency), as well as non-prosodic lexical and discourse features.

Recent work on classifying level of certainty uses pitch and energy features similar to our set of features, plus additional F0 features to better approximate the pitch contour, and non-prosodic word-position features (Litman et al., 2009).

3.5 Chapter Summary

We discussed prosody as a perceptual phenomenon, and characterizations of the pragmatics of prosodic events. We described the set of low-level acoustic pitch, intensity, and temporal features that we computed for our experiments and the correlations, in our corpus, between these features and level of certainty.

The gap between these two levels of prosodic categories—the perceptual level and the acoustic level—is a limitation for the current research on inferring speaker affect. The exploration of more advanced and data-driven methods for prosodic feature selection is an important current direction.

Chapter 4

Modeling Perceived Level of Certainty

We now turn our attention to models for inferring a speaker's *perceived* level of certainty. As discussed in section 2.4.1, we use the term 'perceived level of certainty' to refer to the average of the five annotators' judgments. Knowing the perceived level of certainty is useful in dialogue applications such as tutorial dialogue and voice search. Further, knowing the perceived level of certainty in conjunction with the self-reported level of certainty is useful in educational technology applications, as we discussed in section 1.2.

In our machine learning experiments, we focus on prosodic features as inputs to our models. To see whether this prosodic information complements or overlaps with relevant textual information, we construct a baseline model that takes only non-prosodic features as input. These features are described in section 4.1. We then describe an initial model that uses only basic (utterance-level) prosodic features in section 4.2 and a more sophisticated model that uses contextual prosody features in section 4.3. Some of the work in this chapter was published previously (Pon-Barry and Shieber, 2009a,b, 2011).

4.1 Non-Prosodic Baseline

We want to ensure that the predictions our prosodic models make are not able to be explained by non-prosodic features such as a word's length, familiarity, or part-of-speech, or an utterance's position in the data collection materials. Therefore, as a baseline we construct a set of non-prosodic features.

The features included in this non-prosodic baseline model are listed below ($word_i$ = target word; $word_{i-1}$ = word preceding target word).

- Length in characters of $word_i$
- Number of phonemes in $word_i$
- Number of syllables in $word_i$
- Number of times speaker has previously uttered $word_i$ in experiment
- Frequency of $word_i$ in British National Corpus
- Position of $word_i$ from start of sentence
- Position of $word_i$ from end of sentence
- Relative position of $word_i$ in sentence
- Postion of utterance within experiment
- Part of speech of $word_i$: JJ, NN, NNP
- Part of speech of $word_{i-1}$: DT, PRP$, POS, VBD, CC, RB, NN, TO/IN

There are five types of non-prosodic features: part-of-speech, utterance position, word position, word length, and word familiarity. Nearly all of the features assume knowledge of the sentence's *target word*—the word or phrase that is the locus of uncertainty (see section 2.2.3).

The *part-of-speech* features include binary features for the possible parts-of-speech of the target word and of its immediately preceding word. *Utterance position* is represented as the utterance's ordinal position among the sequence of items in the experiment (varies for each speaker). *Word position* features include the target word's index from the start of the utterance, its index from the end, and its relative position (index from start divided by total words in utterance). The *word length* features include the number of characters, phonemes, and syllables in the target word. To account for *familiarity*, we include a feature for how many times during the experiment the speaker has previously uttered the target word. We also consider a word's general familiarity by looking at word frequency. We use a word's log-probability based on British National Corpus counts to approximate word frequency. For words that do not appear in the British National corpus, we estimate feature values by using web-based counts (Google hits) to interpolate unigram frequencies. Keller and Lapata (2003) demonstrated that using web-based counts is a reliable method for estimating unseen *n*-gram frequencies, see Figure 4.1.

4.2 Basic Prosody Model

The first model that we present is a basic machine learning model that uses only prosodic information to estimate the speaker's level of certainty for a single utterance. We then map the predicted level of certainty score to a categorical classification (certain, uncertain, or

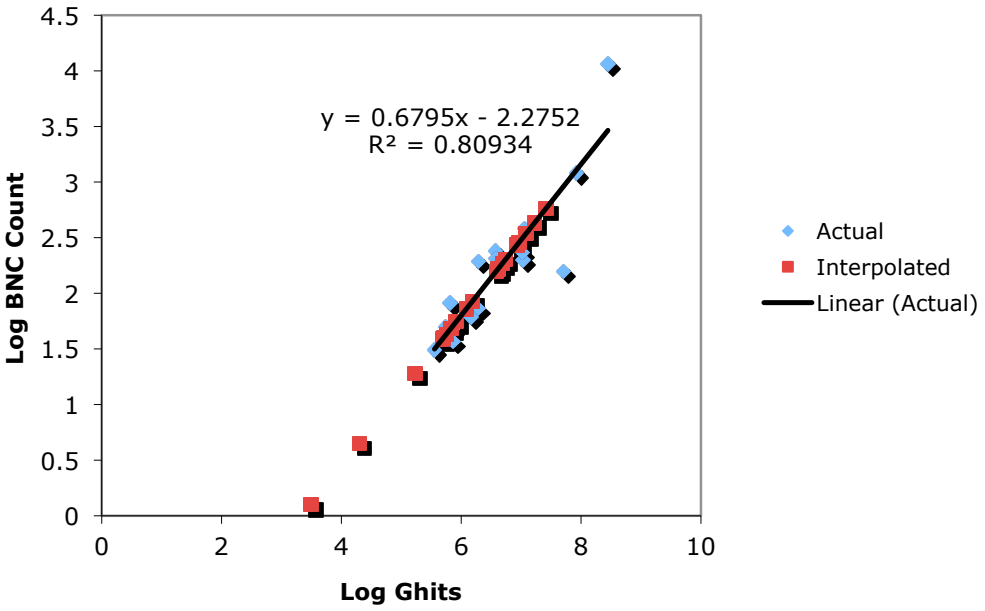

Figure 4.1: Word frequency interpolation.

neutral) to compare our model with prior work. Our basic prosody model performs better than a trivial baseline model (choosing the most common class), corroborating results of prior work, but we also show for the first time that the prosody is crucial in achieving this performance by comparing to a substantive non-prosodic baseline.

4.2.1 Method

The basic prosody model is trained on the 20 utterance-level prosodic features described in section 3.2. We use these features as input variables to a simple linear regression model for predicting perceived level of certainty (on a 1 to 5 scale).

To evaluate our model, we divide the data into 20 folds, one fold per speaker, and perform a k-FOLD CROSS-VALIDATION, where $k = 20$. That is, we fit a model using data

from 19 speakers and test on the remaining speaker. This procedure is repeated for each fold and the results are averaged across all folds. Thus, when we test our models, we are testing the ability to classify utterances of an unseen speaker. We also refer to this procedure as a *leave-one-speaker-out* cross-validation.

We use a simple linear regression model to output real-valued level of certainty predictions. (In section 4.4 we will see why we prefer real-valued output over class-based output.)

We compare our basic prosody model against two baselines: one that always chooses the most common class and one trained on the set of non-prosodic features described in section 4.1. Because the basic prosody model doesn't assume knowledge of the target word, we consider this to be a generous baseline for this experiment. (In section 4.3, we present a prosody model that *does* assume knowledge of the target word.)

4.2.2 Results

Our basic prosody model performs better than the two baseline models and at the same level as previous work for classifying perceived level of certainty.

When comparing root-mean-squared (RMS) error of the basic prosody linear regression model and the non-prosodic linear regression model, we see that the model using prosodic features is a better fit. The basic prosody model has an RMS error of 0.74 whereas the non-prosodic baseline model has an RMS error of 1.06. Table 4.1 shows these results.

The second baseline is to consider the naive model that chooses the most common class. This is the baseline used by Liscombe et al. (2005), whose prior work on level of certainty classification uses decision trees models and outputs one of three classes: certain, uncertain,

Table 4.1: Our basic prosody (linear regression) model has lower RMS error than the non-prosodic baseline model.

Feature Set	**RMS Error**
Non-prosodic features (baseline)	1.06
Utterance-level prosody features	0.74

or neutral. The prosodic features used in our basic prosody model are similar to their set of prosodic features. For their level of certainty classification experiment, Liscombe et al. (2005) use a data set consisting of utterances from an intelligent tutoring system. In their corpus, 'neutral' is the most common class, comprising 66.00% of the utterances.

Our model outputs a real-valued level of certainty score. To compare our model against theirs, we convert our scores into three classes by first rounding to the nearest integer, and then mapping 1 and 2 to uncertain, 3 to neutral, and 4 and 5 to certain. (This partition of the scores is the one that maximizes inter-annotator agreement, see section 2.4.1.) Under this mapping, the most common class in our corpus is 'neutral', comprising 56.25% of the utterances. So for our corpus, the naive baseline of choosing the majority class would yield an accuracy of 56.25%. This is lower than the naive baseline of Liscombe et al. (2005).

Table 4.2 shows the results comparing our basic prosody model against this prior work. Liscombe et al. (2005) compare their model against the naive baseline of choosing the most common class. For their corpus, this baseline was 66.00%. In our corpus, choosing the most-common class gives an accuracy of 56.25%. Our model's classification accuracy is 68.96%, a 29.05% reduction in error compared to choosing the most common class. The comparable model, using a similar set of prosodic features, of Liscombe et al. (2005) has an accuracy of 75.00%, a 26.47% reduction in error compared to choosing the most common

Table 4.2: Our basic prosody model performs significantly better than a linear regression model trained on non-prosodic features, as well as the naive baseline of choosing the most common class. The improvement over this naive baseline is on par with prior work (Liscombe et al., 2005).

Model	**Accuracy**	**Reduction in Error**
Non-prosodic	51.00	
Liscombe most common class	66.00	
Liscombe prosody model	75.00	26%
Our most common class	56.25	
Our basic prosody model	68.96	**29%**

class.

4.2.3 Discussion

Our basic prosody model's improvement over the naive baseline is on par with prior work on certainty classification (Liscombe et al., 2005); while the raw accuracy is lower than the prior work, the reduction in error over the baseline is greater. We consider our evaluation to be more rigorous than that of the prior work. While we test our model using a leave-one-speaker-out cross-validation approach, Liscombe et al. (2005) randomly divide their data into training and test sets. Since they run only a single split, their reported accuracy may not be indicative of the entire data set.

The basic prosody model performs better than a trivial baseline model (choosing the most common class), corroborating results of prior work, but we also show for the first time that the prosody is crucial in achieving this performance by comparing to a substantive

non-prosodic baseline.

4.3 Contextual Prosody Model

In the previous section, we showed that our basic prosody model performs better than two baseline models and on par with previous work. In this section, we show how to improve upon our basic prosody model through context-based feature selection. Because the nature of our corpus makes it possible to isolate a single word or phrase responsible for variations in a speaker's level of certainty (see chapter 2), we have good reason to consider using prosodic features not only at the utterance level, but also at the sub-utterance level (i.e., a single word or a multi-word phrase).

4.3.1 Method

For each utterance, we compute three values for each of the 20 prosodic feature types (see Table 3.1): one value for the whole utterance, one for the context segment, and one for the target word segment, resulting in a total of 60 prosodic features per utterance. While the 20 feature types listed in section 3.2 are comparable to those used in previous uncertainty classification experiments (Liscombe et al., 2005), to our knowledge, no previous work has used features extracted from context or target word segments.

In this experiment, we consider several different sets of prosodic input features. We train linear regression level-of-certainty classifiers and evaluate the models using a leave-one-speaker-out cross-validation procedure, as described in section 4.2. We call the set of 20 whole utterance features from the basic model set A. Set B contains only target word

features. Set C contains only context features. Set D is the union of A, B, and C. And lastly, set E is the 'combination' feature set — a set of 20 features that we designed based on our correlation analysis. For each prosodic feature-type (i.e., each row in Table 3.1) we select either the whole utterance feature value, the context feature value, or the target word feature value, whichever one has the strongest correlation with perceived level of certainty. The features comprising the combination set are listed below.

1. **Whole Utterance:** total silence, total duration, speaking duration, relative position max f0, relative position max RMS, absolute slope (Hz), absolute slope (semitones)

2. **Context:** min f0, max f0, mean f0, stdev f0, range f0, min RMS, max RMS, mean RMS, relative position min RMS

3. **Target Word:** percent silence, speaking rate, relative position min f0, stdev RMS

4.3.2 Results

Table 4.3 shows the accuracies obtained by the linear regression and support vector machine regression models trained on the five subsets of features. The numbers reported are averages of the 20 cross-validation accuracies. To compare these results with those in Table 4.2, we convert the regression output to certain, uncertain, and neutral classes, as described in section 4.2.2. As before, the naive baseline is the accuracy that would be achieved by always choosing the most common class, and the non-prosodic baseline model is the same as described in section 4.1.

Table 4.3: Average classification accuracies for the linear regression and support vector machine models trained on five subsets of prosodic features. The models trained on the *Combination* feature set and the *All* feature set perform better than the other three models in both the 3- and 5-class settings.

Feature Set	**Num Features**	**Accuracy (5 classes)**	**Accuracy (3 classes)**
Naive Baseline	N/A	31.46	56.25
Non-prosodic Baseline	20	29.17	51.00
Linear regression models			
(A) Utterance	20	39.00	68.96
(B) Target Word	20	43.13	68.96
(C) Context	20	37.71	67.50
(D) All	60	48.54	74.58
(E) Combination	20	45.42	74.79
Support vector machine models			
(A) Utterance	20	39.60	69.85
(B) Target Word	20	38.80	68.10
(C) Context	20	36.05	63.65
(D) All	60	38.75	70.20
(E) Combination	20	40.45	70.20

4.3.3 Discussion

The key comparison to notice is that the combination feature set E, with only 20 features, yields higher average accuracies than the utterance feature set A: a difference of 5.83%. This suggests that using a combination of features from the context and target word in addition to features from the whole utterance leads to a better model of the perceived level of certainty than using features from only the whole utterance.

Each fold in our cross-validation corresponds to a different speaker, so the folds are *not* identically distributed and we do not expect each fold to yield the same prediction accuracy. That means that we should compare predictions of the two feature sets within folds rather than between folds. Figure 4.2 shows the correlations between the predicted and perceived levels of certainty for the models trained on sets A and E. The combination set E predictions were more strongly correlated than whole utterance set A predictions in 16 out of 20 folds. This result supports our claim that using a combination of features from the context and target word in addition to features from the whole utterance leads to better prediction of level of certainty.

Figure 4.2 also shows that one speaker (the 17th fold) is an outlier—for this speaker, our model's level of certainty predictions are less correlated with the perceived levels of certainty than for all other speakers. Most likely, this results from non-prosodic cues of uncertainty present in the utterances of this speaker (e.g., disfluencies). Removing this speaker from our training data did not improve the overall performance of our models.

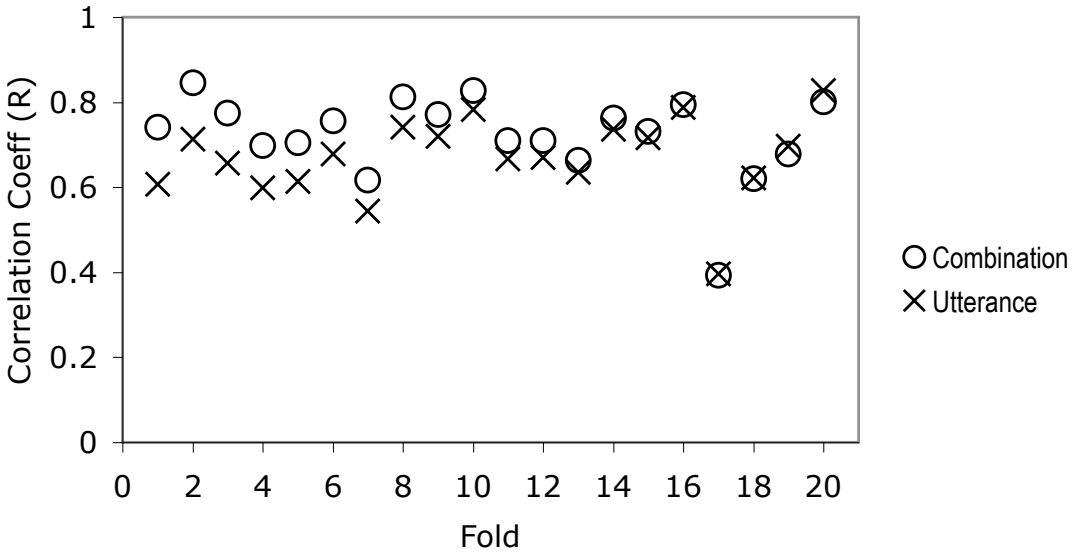

Figure 4.2: Correlations with perceived level of certainty per fold for the *Combination (O)* and the *Utterance (X)* feature set predictions, sorted by the size of the difference.

4.4 Determining the Locus of Uncertainty

In the previous section, 4.3, we presented machine learning models that predict level of certainty for an utterance, given a vector of prosodic features as input. The models did better when the locus of uncertainty was known. In this section, we show that even when we have no a priori knowledge of the locus of uncertainty, our models can be used to make an informed guess about the underlying locus of uncertainty. As an initial step, we consider the problem of selecting one phrase, among two candidate phrases, as the underlying locus of uncertainty. In other words, if we compare the prosody of two phrases, one that the speaker is uncertain about (the actual target word), and another phrase that he or she is certain about (a control word), can our models determine which phrase is the cause of uncertainty? The subsequent evaluation would be to consider all possible words and phrases.

Our approach is to use the level-of-certainty regression models described in section 4.3. We compare the level of certainty scores generated by these models when using prosodic features relative to the actual target word versus using prosodic features relative to the control word. Our best model is able to identify the correct target word 91% of the time, a 71% error reduction over a baseline model that is trained on only non-prosodic features.[1]

4.4.1 Method

For a subset of utterances that were perceived to be uncertain (average level of certainty less than 2.5), we identify a control word — a content word roughly the same length as the potential target words and if possible, the same part-of-speech. In the example item shown below, the control word is *abrasive*.

(Ex. 3) `Mahler's revolutionary music, abrasive personality, and _____ writings about art and life divided the city into warring factions.`

a. `officious`

b. `trenchant`

c. `spoffish`

d. `pugnacious`

We balance the set of control words for position in the utterance relative to the position of the actual target word; half of the control words appear before the actual target word and half appear after. After filtering utterances based on level of certainty and presence of an appropriate control word, 43 utterances remain. This is the test set for this experiment.

[1] Some of the results in this chapter have been previously published (Pon-Barry and Shieber, 2009b).

We then extract vectors of prosodic features for two segmentations of the utterance: (a) the correct segmentation with the actual target word as the proposed 'target word' and (b) an alternative segmentation with the control word as the proposed 'target word.' Thus, the prosodic features extracted from the target word and from the context will be different in these two segmentations, while the features extracted from the utterance will be the same. We then compare the predicted levels of certainty of the two candidates. The hypothesis we test in this experiment is that our models should predict a lower level of certainty when the prosodic features are taken from segmentation (a) rather than segmentation (b), thereby identifying the actual target word as the source of the speaker's uncertainty.

The procedure is illustrated in Figure 4.3. For each $utterance_i$, we compute the vector of prosodic features for each of the two candidate segmentations. We then plug these feature vectors, $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$, into our level of certainty classifiers, which were described in section 4.3. As in the experiments described in section 4.3, this evaluation follows a leave-one-speaker-out cross-validation procedure. We use the non-prosodic model described in section 4.1 as a baseline for this experiment.

4.4.2 Results

On the task of identifying the locus of uncertainty (i.e., the actual target word rather than a control word), we obtain accuracies ranging from 49% to 91% with the linear regression models, and 28% to 81% with the SVM regression models. The wide range in the accuracies depends on the set of features used to train the level of certainty models. Table 4.4 shows the complete set of linear regression and support vector machine accuracies for each set of prosodic and non-prosodic features. The models trained on the non-prosodic

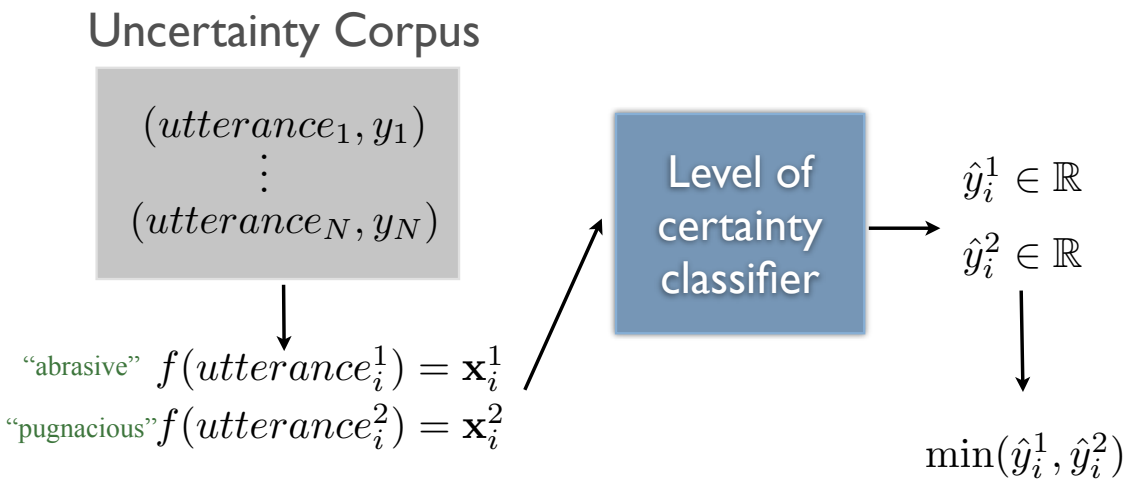

Figure 4.3: Procedure for identifying locus of uncertainty by comparing the predicted certainty scores of the candidate target words.

features provide a baseline from which to compare the performance of the models trained on prosodic features. This baseline accuracy is 67%. The linear regression model that was trained on the set of target word features and non-prosodic features achieves 91% accuracy.

The linear regression model trained on the target word feature set had the highest accuracy among the purely prosodic models, 86%. The highest overall accuracy, 91%, was achieved on the model trained on the target word features plus the non-prosodic features from the baseline set.

4.4.3 Discussion

This experiment shows that prosodic level-of-certainty models are useful in detecting uncertainty at the word level. Our best model, the one that uses prosodic target word features plus the non-prosodic features from the baseline set, identifies the correct word 91% of the time whereas the baseline model using only non-prosodic features is accurate

Table 4.4: Accuracies on the task of identifying the locus of uncertainty, when choosing between the actual target word and a control word.

Feature Set	**Number of Features**	**Linear Regression**	**Support Vector Machine**
Non-prosodic (baseline)	20	67.44%	67.44%
Target Word, Non-prosodic	40	90.70%	79.07%
Target Word	20	86.05%	79.07%
Target Word, Context, Utterance	60	79.07%	81.40%
Target Word, Context, Utterance, Non-prosodic	80	76.74%	81.40%
Target Word, Utterance	40	69.77%	72.09%
Combination Set: Target Word	4	72.09%	67.44%
Combination Set	20	72.09%	53.49%
Context	20	48.84%	27.91%

just 67% of the time. This is an absolute difference of 23% and an error reduction of 71%. This large improvement over the non-prosodic baseline model implies that target word prosodic features are crucial in word-level uncertainty detection.

In creating the non-prosodic feature set for this experiment we wanted to account for the most obvious differences between the target words and the control words. The baseline model's low accuracy on this task is to be expected because the non-prosodic features are not good at explaining the variance in the response variable (perceived level of certainty): the correlation coefficient for the baseline linear regression model is only 0.27. As a comparison, the coefficient for the target word linear regression model is 0.67.

The combination feature set, which had high accuracy in classifying an utterance's overall level of certainty, did not perform as well as the other feature sets for this word-level task. We speculate that this may have to do with the context features. While the prosodic features we extracted from the context may be beneficial in classifying an utterance's overall level of certainty, the low accuracies for the context feature set on this task suggest that they are detrimental in determining which word a speaker is uncertain about. The task we examine in this section, distinguishing the actual target word from a control word, is different from the task the models are trained on (predicting a real-valued level of certainty); therefore we do not expect the models with the highest classification accuracy to necessarily perform well on the task of identifying the word causing uncertainty.

It is not clear which model type, linear regression or support vector machine regression, is better suited for the general task of identifying uncertain words within an utterance. Among all the models we trained, the two with the highest accuracies were both linear regression models (one with only prosodic features and one with a mixture of prosodic

and non-prosodic features). However, the SVM models yielded higher accuracies than the linear regression models for three of the eight sets of features examined.

4.5 Chapter Summary

In this chapter, we presented a basic prosody model and a contextual prosody model. The basic prosody model, which uses utterance-level prosodic features as input, performs better than a naive baseline model (choosing the majority class) and better than a substantive baseline model that uses non-prosodic features. This basic prosody model achieves a classification accuracy that is on par with the classification results of prior work. The contextual prosody model, which uses sub-utterance prosodic features achieves an even greater classification accuracy of 75%.

These results suggest a better predictive model of level of certainty for systems where words or phrases likely to cause uncertainty are known ahead of time, for example in a tutorial dialogue system or language learning system (where the system is taking the initiative). Without increasing the total number of features, combining select prosodic features from the target word, the surrounding context and the whole utterance leads to better prediction of level of certainty than using features from the whole utterance only.

In section 4.4, we saw that these same models can be useful even when the words likely to cause uncertainty are *not* known ahead of time. If we have an utterance where we believe the speaker is uncertain, but we do not know where the locus of uncertainty is, the level of certainty regression models from section 4.3 can be used to find the locus of uncertainty.

Chapter 5

Modeling Internal Level of Certainty

In chapter 4, we focused on classifying a speaker's perceived level of certainty. Here, we turn our attention to modeling a speaker's internal level of certainty. Our examination of internal affect is a key way in which this work goes beyond existing work on affect classification. To our knowledge, no other studies have attempted to control a person's *internal* affective state.

In the data collection procedure described in chapter 2, we obtained self-reported certainty ratings from the participants. We showed that the perception of certainty does not necessarily match the self-reported levels. In this chapter we look in more detail at two issues regarding internal level of certainty: using prosodic information to predict self-reported certainty, and assessing how well self-reports reflect a speaker's actual internal state. First, in section 5.1, we present experiments on classifying self-reported levels of certainty using low-level prosodic features (the same features that were outlined in chapter 3). In section 5.2, we show that level of certainty self-reports are a good proxy for internal level of certainty, for the handwritten digit domain. We find that internal and self-

reported certainty are highly correlated ($r = -0.818$). This result is novel; the comparison of internal versus self-reported affect has not received significant attention in the affect detection community.

5.1 Modeling Self-Reports in the Uncertainty Corpus

In this section, we present our experiments aimed at classifying a person's self-reported level of certainty. In our initial experiments, we train models on only prosodic features of the utterance. We then consider the possibility of using information gleaned from perceived level of certainty to more accurately model the self-reported level. This idea bears promise especially given the potential, pursued by ourselves (see chapter 4) and others (Liscombe et al., 2005), of inferring the perceived level of certainty directly from prosodic information. We show that a kind of triage on the perceived level of certainty can improve self-report predictions.

5.1.1 Basic Prosodic Feature Set

As an initial model, we train a single decision tree using the 20 prosodic features listed in section 3.2. We use C4.5 decision tree models, with the Weka (Hall et al., 2010) toolkit. We use a 20-fold leave-one-speaker-out cross-validation approach (as described in section 4.2) to evaluate this model over all the utterances in the Uncertainty Corpus.

Results

This initial decision tree model, that uses only prosodic features, classifies self-reports with an accuracy of 66.33%. As shown in Table 5.1, the single decision tree model does

Table 5.1: Accuracy in classifying self-reported level of certainty, for the initial prosody decision tree model and for two baseline models.

Model	**Accuracy**
Baseline 1: Majority Class	52.30
Baseline 2: Assign Perceived Level	63.67
Single Prosody Decision Tree	66.33

better than the naive baseline of choosing the most-common class, which has an accuracy of 52.30%, and marginally better than assigning the self-reported certainty to be the same as the perceived certainty, which has an accuracy of 63.67%. Still, we would like to know if we could do better than 66.33%.

5.1.2 Extended Feature Set

As an alternative approach, suppose we know an utterance's perceived level of certainty. Could we use this knowledge, along with the prosody of the utterance to better predict the self-reported certainty?

To test this, we divide the data into four subsets (see Figure 5.1) corresponding to the correctness of the answer and the perceived level of certainty.

As illustrated in figure 5.2, the distribution of self-reports in subset A', is heavily skewed: 84% of the utterances are self-reported as *uncertain*. This imbalance is intuitive; someone who is incorrect *and* perceived as uncertain most likely feels uncertain too. Likewise, in subset B', the distribution of self-reports is skewed in the other direction: 76% of the utterances in this subset are self-reported as *certain*. This too is intuitive; someone

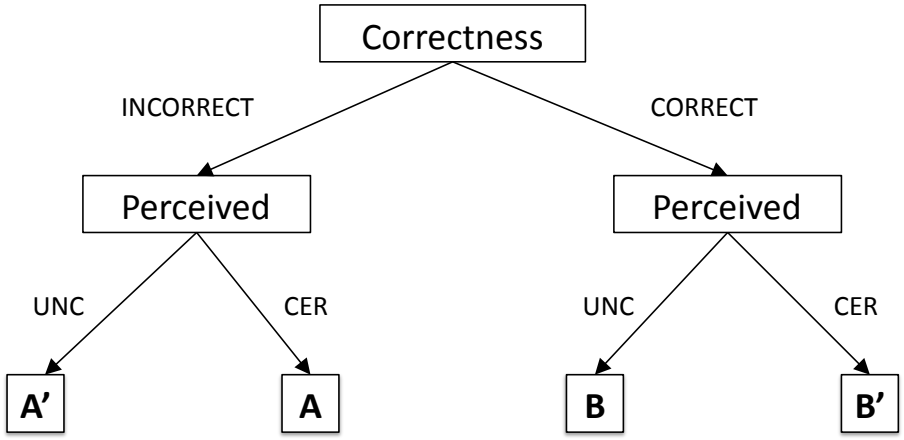

Figure 5.1: Division of utterances into four subsets, based on answer correctness and perceived certainty.

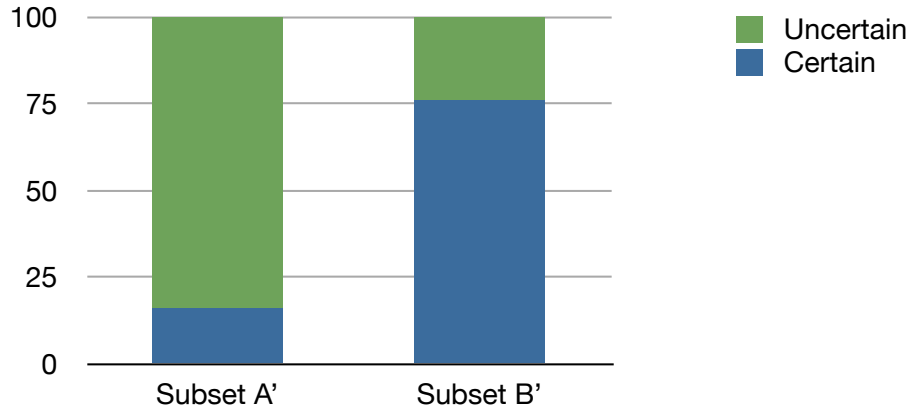

Figure 5.2: Distribution of self-reports in subsets A' and B'.

who is correct *and* perceived as certain most likely feels certain as well. Therefore, we hypothesize that for subsets A' and B', classification models trained on prosodic features will do no better than choosing the subset-specific majority class.

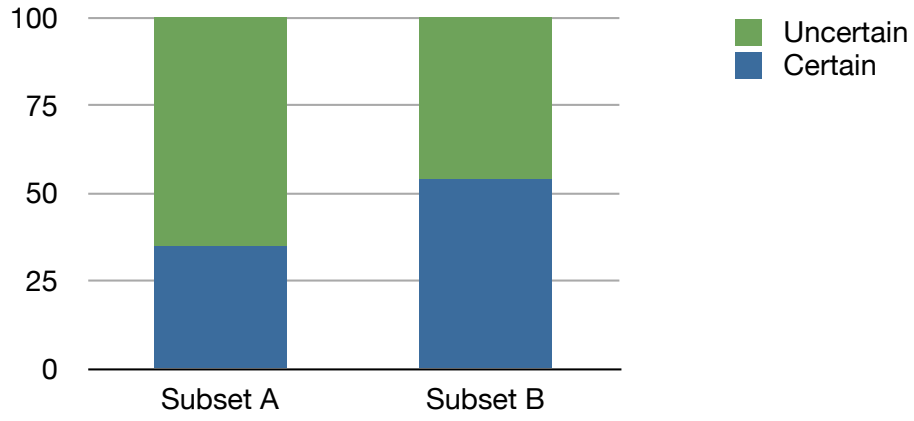

Figure 5.3: Distribution of self-reports in subsets A and B.

Subsets A and B are the more interesting cases; they are the subsets where the perceived level of certainty is not aligned with the correctness. As illustrated in figure 5.3, the self-reported levels of certainty for these subsets are less skewed: 65% *uncertain* for subset A and 54% *certain* for subset B. We hypothesize that for subsets A and B, decision tree models trained on prosodic features will be more accurate than selecting the subset-specific majority class. For each subset, we perform a k-fold cross-validation where we leave one speaker out of each fold. Because not all speakers have utterances in every subset, k ranges from 18 to 20.

Results

For subset A, the decision tree accuracy in classifying the self-reported level of certainty is 68.99%, while assigning the subset-majority class (uncertain) results in an accuracy of

Table 5.2: Accuracies for classifying self-reported level of certainty for the prosodic decision tree models trained separately on each of the four subsets of utterances.

Subset	**Accuracy (Subset Majority)**	**Accuracy (Prosody Decision Tree)**
A	65.19	**68.99**
B	53.52	**69.01**
A′	84.35	84.35
B′	75.89	75.89
Overall	72.49	**75.30**

65.19%. For subset *B*, the decision tree accuracy is 69.01%, while assigning the subset-majority class (certain) results in an accuracy of 53.52%. Thus, for these two subsets, the prosody of the utterance is more informative than the majority class baseline. As expected, for subsets A' and B', the decision tree models do no better than assigning the subset-majority class. These results are summarized in Table 5.2.

The combined decision tree model has an overall accuracy of 75.30%, significantly better than the accuracy of single decision tree model (66.33%, see Table 5.1), which assumed no knowledge of the correctness or the perceived level of certainty. Therefore, if we know an utterance's perceived level of certainty, we can use that information to much more accurately model the self-reported level of certainty.

Our combined decision tree model also outperforms the decision tree that has knowledge of prosody and of correctness but no knowledge of perceived certainty; this tree ignores the prosody and splits only on correctness, which was equivalent to choosing the

subset majority class (72.49% accuracy).

5.1.3 Discussion

In our decision tree models we find that the *percent silence* and *speaking rate* features are consistently selected[1] as attributes to 'split' on, in other words, they lead to the highest information gain. Lower values of percent silence correspond to speakers feeling certain and higher values correspond to speakers feeling uncertain. For speaking rate, very slow and very fast speaking rates correspond to speakers feeling certain. Values in the middle correspond to a mix of speakers feeling certain and uncertain. In our correlation analysis (see section 3.3), speaking rate was not strongly correlated with perceived level of certainty. This suggests that perhaps speaking rate is important in distinguishing internal levels of certainty from perceived levels of certainty.

We find that classifying the speaker's self-reported level of certainty, with knowledge of the utterance's correctness and perceived level of certainty, is a better way to learn the speaker's self-awareness and transparency than classifying self-awareness and transparency directly. Still, classifying the speaker's self-reported level of certainty is a difficult problem. Our decision tree model results suggest that prosodic information is helpful in this classification task. In addition, our results suggest that *speaking rate* may be useful in distinguishing self-reported ratings from perceived ratings.

[1]Speaking rate was one of the first two attributes split on in 100% of the cross-validation trials; percent silence was one of the first two attributes in 95% of the cross-validation trials.

5.2 Handwritten Digit Experiment Results

For the handwritten digit domain, we elicit self-reported levels of certainty and we control each item's intrinsic level of certain, as described in section 2.3. Our analysis comparing these two quantities indicates that self-reported certainty is a good proxy for internal level of certainty. This is illustrated in Figure 5.4, where each point corresponds to one of the handwritten digit images. The y-axis value is the mean self-reported certainty, averaged over all participants. The x-axis value is the entropy of the human-generated label distribution, based on the Mechanical Turk data collection (see section 2.3.1). The correlation coefficient of these two quantities is -0.818. When the self-reports are standardized by speaker first (instead of using the raw values), we get a similar result, with a correlation coefficient of -0.817.

5.3 Chapter Summary

In section 5.1, we presented experiments on classifying level of certainty using low-level prosodic features and knowledge of the perceived level of certainty. In section 5.2, we present correlation results suggesting that self-reports are indeed an accurate proxy for internal level of certainty.

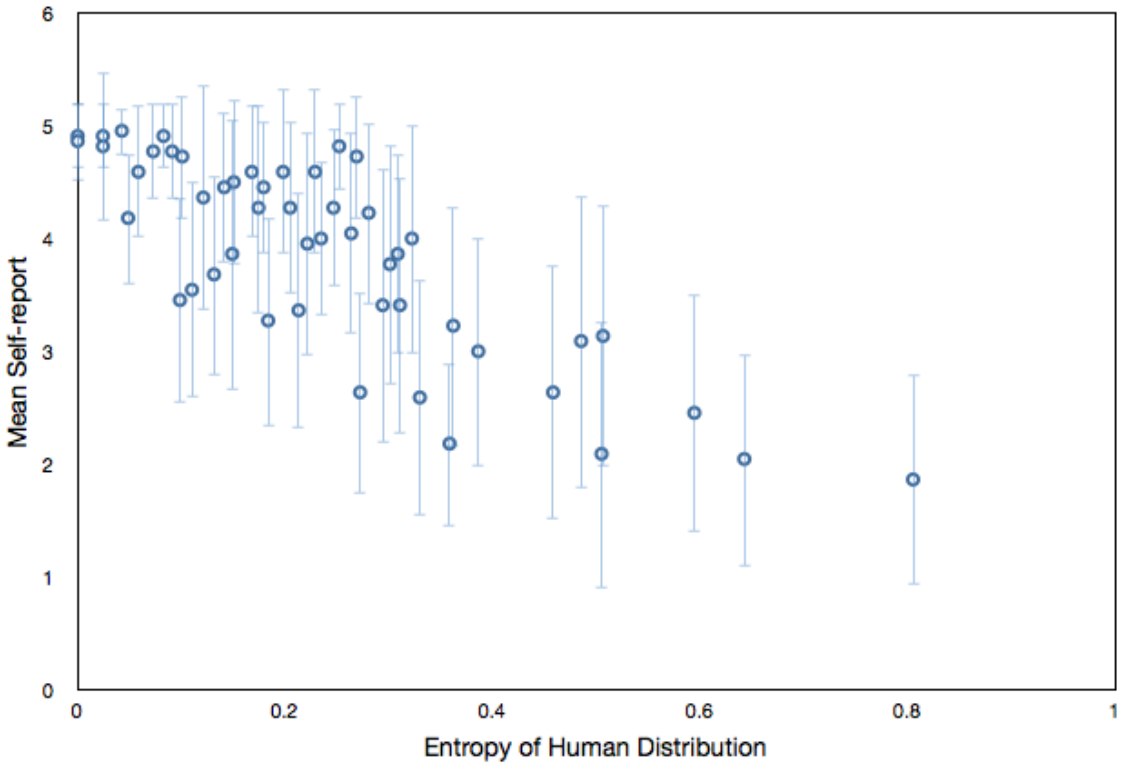

Figure 5.4: Self-reported certainty versus entropy.

Chapter 6

Conclusion

The results presented in this thesis suggest that we can get a good estimate of a speaker's level of certainty based on only prosodic features. Further, they indicate that knowledge of which phrases were likely to have engendered uncertainty can significantly enhance the system's ability to predict level of certainty, and even to select which phrase is the source of uncertainty.

In our experiments, we used a small set of the many possible prosodic features that have been examined in related work. Because these features proved beneficial in recognizing uncertainty, we believe that using an expanded set of prosodic features might be even more beneficial. In natural conversation, people also convey uncertainty through other channels such as body language, facial gestures, and word choice. Further work is needed to understand how to integrate cues from multiple modalities, when these other modes of input are available.

Our results were enabled by a novel methodology for collecting uncertainty data that allowed us to isolate the phrase causing uncertainty. We also addressed a question that is

important to all research regarding mental or emotional state modeling — the difference between a person's internal state and an outsider's perception of that state. In our corpus, these two quantities are aligned for approximately one-half of the utterances and mismatched for the remaining half, suggesting that classifiers trained on only perceived judgements of certainty may end up missing actual instances of uncertainty. This highlights the importance of collecting data in ways that maximize our ability to externally control or ensure access to a person's internal mental state. It also raises the question of whether computers may even surpass humans at classifying a speaker's internal level of certainty.

Bibliography

Jaime Acosta and Nigel Ward. Responding to user emotional state by adding emotional coloring to utterances. In *Proceedings of Interspeech 2009*, pages 1587–1590, Brighton, UK, 2009.

Abeer Alwan, Yijian Bai, Matt Black, Larry Casey, Matteo Gerosa, Margaret Heritage, Markus Iseli, Barbara Jones, Abe Kazemzadeh, Sungbok Lee, Shrikanth Narayanan, Patti Price, Joseph Tepperman, and Shizhen Wang. A system for technology based assessment of language and literacy in young children: the role of multiple information sources. In *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, pages 26–30, Chania, Greece, 2007.

Shankar Ananthakrishnan and Shrikanth Narayanan. Automatic prosody. *IEEE Transactions on Speech and Audio Processing*, 2008.

Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2037–2040, Denver, CO, 2002.

Matthew P. Black, Panayiotis G. Georgiou, Athanasios Katsamanis, Brian R. Baucom, and Shrikanth S. Narayanan. "You made me do it": Classification of blame in married couples' interactions by fusing automatically derived speech and language information. In *Proceedings of Interspeech*, Florence, Italy, 2011.

Paul Boersma and David Weenink. Praat: doing phonetics by computer [version 5.0.19], January 2010. URL `http://www.praat.org/`.

Dwight Bolinger. *Intonation and its Uses*. Stanford University Press, Stanford, CA, 1989.

Susan E. Brennan and Maurice Williams. The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34:383–398, 1995.

Josefa Heifetz Byrne. *Mrs. Byrne's dictionary of unusual, obscure, and preposterous words: gathered from numerous and diverse authoritative sources*. University Books, Secaucus, NJ, 1974.

Chris Callison-Burch and Mark Dredze. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics, 2010.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991. ISBN 0-471-06259-6.

Roddy Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80, January 2001. ISSN 1053-5888. doi: 10.1109/79.911197.

Roddy Cowie, Naomi Sussman, and Aaron Ben-Ze'ev. Emotion: Concepts and definitions. In Paolo Petta, Pelachaud Catherine, and Roddy Cowie, editors, *Emotion-Oriented Systems*. Springer, Berlin Heidelberg, 2010.

Scotty D. Craig, Arthur C. Graesser, Jeremiah Sullins, and Barry Gholson. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media*, 29(3):241–250, 2004.

Raul Fernandez. *A Computational Model for the Automatic Recognition of Affect in Speech*. PhD thesis, Massachusetts Institute of Technology, 2004.

Kate Forbes-Riley and Diane Litman. Adapting to student uncertainty improves tutoring dialogues. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, pages 33–40, Brighton, UK, 2009.

Kate Forbes-Riley, Diane Litman, and Mihai Rotaru. Responding to student uncertainty during computer tutoring: an experimental evaluation. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, pages 60–69, Montreal, Canada, 2008.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP.*, volume 1, pages 517–520. IEEE, 1992.

Barbara Grosz and Candice Sidner. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. WEKA data mining software [version 3.6.2], January 2010. URL `http://www.cs.waikato.ac.nz/ml/weka/`.

Julia Hirschberg. Intonation and pragmatics. In L. Horn and G. Ward, editors, *Handbook of Pragmatics*. Blackwell, 2003.

Julia Hirschberg, Diane Litman, and Marc Swerts. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43(1–2):155–175, 2004.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67. ACM, 2010.

M. Iseli, Y.-L. Shue, M. Epstein, P. Keating, J. Kreiman, and Abeer Alwan. Voice source correlates of prosodic features in american english: a pilot study. In *Proceedings of Interspeech*, pages 2226–2229, Pittburgh, PA, 2006.

Frank Keller and Mirella Lapata. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484, 2003.

Emiel Krahmer and Marc Swerts. How children and adults produce and perceive uncertainty in audiovisual speech. *Language and Speech*, 48(1):29–53, 2005.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

Chul Min Lee and Shrikanth Narayanan. Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, 2005.

Gina-Anne Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings of COLING-ACL*, pages 736–742, 1998.

Gina-Anne Levow. Adaptations in spoken corrections: Implications for models of conversational speech. *Speech Communication*, 36(1–2):147–163, 2002.

Gina-Anne Levow. Unsupervised and semi-supervised learning of tone and pitch accent. In *Proceedings of HLT-NAACL*, pages 224–231, 2006.

Jackson Liscombe, Julia Hirschberg, and Jennifer Venditti. Detecting certainness in spoken tutorial dialogues. In *Proceedings of Eurospeech*, Lisbon, Portugal, 2005.

Jackson J. Liscombe. *Prosody and speaker state: paralinguistics, pragmatics, and proficiency*. PhD thesis, Columbia University, 2007.

Diane Litman and Kate Forbes-Riley. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 13:559–590, 2006.

Diane Litman and Kate Forbes-Riley. Spoken tutorial dialogue and the feeling of another's knowing. In *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 286–289, London, UK, 2009.

Diane Litman and Scott Silliman. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proceedings of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, Boston, MA, 2004.

Diane Litman, Mihai Rotaru, and Greg Nicholas. Classifying turn-level uncertainty using word-level prosody. In *Proceedings of Interspeech*, pages 2003–2006, Brighton, UK, 2009.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

Subhransu Maji and Jitendra Malik. Fast and accurate digit classification. Technical Report UCB/EECS-2009-159, EECS Department, University of California, Berkeley, November 2009. URL `http://ttic.uchicago.edu/~smaji/projects/digits/`.

Winter Mason and Siddharth Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44:1–23, 2011.

Christine Nakatani, Julia Hirschberg, and Barbara Grosz. Discourse structure in spoken language: Studies on speech corpora. In *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

Christine Hisayo Nakatani. *The computational processing of intonational prominence: A functional prosody perspective*. PhD thesis, Harvard University, 1997.

Tim Paek and Yun-Cheng Ju. Accommodating explicit user expressions of uncertainty in voice search or something like that. In *Proceedings of Interspeech*, pages 1165–1168, Brisbane, Australia, September 2008.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419, August 2010.

Brian Parkinson. Emotions in direct and remote social interaction: Getting through the spaces between us. *Computers in Human Behavior*, 24(4):1510–1529, 2008.

Jonathan W. Pierce. PsychoPy—Psychophysics software in python. *Journal of Neuroscience Methods*, 162(1–2):8–13, 2007.

Jonathan W. Pierce. PsychoPy: Psychology software in python [version 1.74.00], July 2012. URL `http://www.psychopy.org/`.

Janet Pierrehumbert and Julia Hirschberg. The meaning of intonational contours in the interpretation of discourse. In Jerry Cohen, Philip Morgan, and Martha Pollack, editors, *Intentions in Communication*, pages 271–311. Bradford Books (MIT Press), Cambridge, MA, 1990.

Heather Pon-Barry. Prosodic manifestations of confidence and uncertainty in spoken language. In *Proceedings of Interspeech*, pages 74–77, Brisbane, Australia, September 2008.

Heather Pon-Barry and Stuart Shieber. The importance of sub-utterance prosody in predicting level of certainty. In *Proceedings of NAACL-HLT*, Boulder, CO, June 2009a.

Heather Pon-Barry and Stuart Shieber. Identifying uncertain words within an utterance via prosodic features. In *Proceedings of Interspeech*, pages 1579–1582, Brighton, UK, September 2009b.

Heather Pon-Barry and Stuart Shieber. Assessing self-awareness and transparency when classifying a speaker's level of certainty. In *Proceedings of Speech Prosody*, Chicago, IL, May 2010.

Heather Pon-Barry and Stuart M. Shieber. Responding to uncertainty in speech. *EURASIP Journal on Advances in Signal Processing*, 2011.

Heather Pon-Barry, Karl Schultz, Elizabeth Bratt, Brady Clark, and Stanley Peters. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education*, 16:171–194, 2006a.

Heather Pon-Barry, Fuliang Weng, and Sebastian Varges. Evaluation of content presentation strategies for an in-car spoken dialogue system. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech/ICSLP)*, pages 1930–1933, 2006b.

Rajesh Ranganath, Dan Jurafsky, and Dan McFarland. It's not you it's me: Detecting flirting and its misperception in speed-dates. In *Proceedings of EMNLP*, pages 334–342, Singapore, 2009.

Rajesh Ranganath, Dan Jurafsky, and Daniel A. McFarland. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech and Language*, 2012.

Doug Rohde. Linger: a flexible platform for language processing experiments [version 2.94], February 2008. URL `http://tedlab.mit.edu/~dr/Linger/`.

Andrew Rosenberg. *Automatic Detection and Classification of Prosodic Events*. PhD thesis, Columbia University, 2009.

Marc Schröder and Roddy Cowie. Developing a consistent view on emotion-oriented computing. *Machine Learning for Multimodal Interaction*, pages 194–205, 2006.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. The Interspeech 2010 paralinguistic challenge. In *Proceedings of Interspeech*, pages 2794–2797, 2010.

Björn Schuller, Anton Batliner, Stefan Steidl, Florian Schiel, and Jarek Krajewski. The Interspeech 2011 speaker state challenge. In *Proceedings of Interspeech*, pages 3201–3204, 2011.

Stefanie Shattuck-Hufnagel and Alice E. Turk. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25:193–247, 1996.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. TOBI: A standard for labeling english prosody. In *Proceedings of International Conference on Spoken Language Processing*, 1992.

Kåre Sjölander and Jonas Beskow. Wavesurfer — an open source speech tool. In *Proceedings of ICSLP 2000, 6th International Conference on Spoken Language Processing*, pages 464–467, Beijing, 2000.

Kåre Sjölander and Jonas Beskow. Wavesurfer: an open source tool for sound visualization and manipulation [version 1.8.5], March 2008. URL `http://sourceforge.net/projects/wavesurfer/`.

Vicki L. Smith and Herbert H. Clark. On the course of answering questions. *Journal of Memory and Language*, 32:25–38, 1993.

Stefan Steidl, Michael Levit, Anton Batliner, Elmar Nöth, and Heinrich Niemann. "Of all things the measure is man" Automatic classification of emotions and inter-labeler consistency. In *Proceedings of ICASSP*, 2005.

Marc Swerts and Emiel Krahmer. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1):81–94, 2005.

Joseph Tepperman, David Traum, and Shrikanth Narayanan. Yeah right: Sarcasm recognition for spoken dialogue systems. In *Proceedings of InterSpeech*, Pittsburgh, PA, 2006.

Kurt VanLehn, Stephanie Siler, Charles Murray, Takashi Yamauchi, and William B. Baggett. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3):209–249, 2003.

Ye-Yi Wang, Dong Yu, Yun-Cheng Ju, and Alex Acero. Voice search. In Gokhan Tur and Renato DeMori, editors, *Spoken Language Understanding*. Wiley, 2011.

Appendix A

Data Collection Materials

A.1 Phase 1: Instructions to Participants

Read aloud by experimenter:
Pretend that I have just moved to the Boston area. I will ask you a series of questions about using public transportation in Boston. Please answer as best you can, you are not expected to know the answers to all the questions.

A.2 Phase 1: Instructions to Annotators

You are going to be asked to listen to (and categorize) a corpus of 600 utterances that were collected in the following manner: speakers were presented with a partial written sentence and asked to complete the sentence according the some criteria. Some sentences have one gap and some have more than one gap. Speakers were told to read the complete sentence aloud upon hearing a beep. The recordings start immediately after the beep and end when the speaker finishes reading the sentence.

We gave speakers multiple options for completing the gaps. For each utterance, please judge how certain you feel the speaker sounds. Use the following scale:

IMPORTANT: You are not asked to judge how sensible the sentence is, just judge how certain the speaker sounds. Further, you do not need to think about the source of the certainty or uncertainty; make your judgments with respect to the whole utterance.

You may listen to each utterance as many times as you like. However, you should not go back and change previous ratings you have made.

The corpus is divided into 12 sections. We expect that each section will take 20-30 minutes. We will specify the order in which you should work through the sections. In each section, listen to the speech files in the order they are listed. You must complete each section in a single sitting, and do not do more than two sections without taking a break.

A.3 Phase 1: Transit Items

(T-01) Question: From the Harvard T stop, how can I get to the Silver line?

Answer: Take the red line to _____ .

a. South Station

b. Downtown Crossing

(T-02) Question: What is the best way to get to North Station from the Harvard T stop?

Answer: Take the red line to _____

a. Park Station

b. Downtown Crossing

and transfer to the _____ .

a. green line

b. orange line

(T-03) Question: From the Harvard T stop, what is the best way to get to

Chinatown without transferring?

Answer: Take the red line to _____ .

a. South Station

b. Downtown Crossing

(T-04) Question: What is the best way to get from Harvard to the Boston Garden?

Answer: Take the red line to the _____ and get off at North Station.

a. orange line

b. green line

(T-05) Question: How can I get from Harvard to Faneuil Hall on the T?

Answer: Take the red line to the _____

a. green line

b. orange line

and get off at _____ .

c. Haymarket

d. Government Center

(T-06) Question: From Park Station, how can I get to Boston University?

Answer: Take the _____ outbound.

a. green line

b. red line

(T-07) Question: What is the best way to go from Kenmore Square to

South Station?

Answer: Take the green line _____ and

a. inbound

b. outbound

transfer to the _____

c. blue line

d. red line

at _____ and get off at South Station.

e. Park Station

f. Government Center

(T-08) Question: I just watched a Red Sox game at Fenway Park, what is the fastest way to get to the airport?

Answer: Take the green line inbound to the _____

a. red line

b. blue line

then follow signs to the _____ .

c. airport shuttle

a.silver line

(T-09) Question: You are meeting a visitor at South Station and they want to see the aquarium. How do you get there?

Answer: Take the red line to Downtown Crossing, transfer to the _____

a. orange line

b. green line

and then switch to the blue line at _____ .

c. State Street

d. Government Center

(T-10) Question: You are at South Station and you want to go to Porter Square. You learn that no red line trains are running between Alewife and Park Street due to a derailment. You decide to try to get to Porter via the commuter rail. How do you get there?

Answer: From South Station, take the red line inbound to the _____

a. green line

b. orange line

then get off at _____

c. North Station

d. Government Center

and catch a commuter rail train bound for _____ .

e. Fitchburg

f. Lowell

A.4 Phase 1: Vocabulary Items

(V-01) She is a _____ opponent; you must respect and fear her at all times.

a. redoubtable

b. pugnacious

c. craven

d. vituline

(V-02) Your _____ tactics may compel me to cancel the contract.

a. redoubtable

b. dilatory

c. hidrotic

d. sycophantic

(V-03) Lillian's _____ refusal to join the protest was criticized by her comrades.

a. vituline

b. dilatory

c. craven

d. disingenuous

(V-04) If you carry this _____ attitude to the conference, you will alienate any supporters you may have.

a. truculent

b. conciliatory

c. sycophantic

d. hidrotic

(V-05) Fed up by the brownnosers who made up his entourage, the star cried, ''Get out, all of you! I'm sick of your _____ ways!''

a. truculent

b. conciliatory

c. craven

d. spoffish

(V-06) She was still angry despite his _____ words.

a. redoubtable

b. conciliatory

c. sycophantic

d. officious

(V-07) He decided to protest the allegations with the help of the _____ newspaper columnist.

a. redoubtable

b. craven

c. dilatory

d. truculent

(V-08) Only the _____ workers in the office laughed at all of the manager's bad jokes.

a. pugnacious

b. craven

c. sycophantic

d. spoffish

(V-09) I am afraid of his _____ wit for it is so often sarcastic.

a. trenchant

b. disingenuous

c. redoubtable

d. hidrotic

(V-10) The man's _____ clothes and old-fashioned language marked him as an eccentric.

a. spoffish

b. disingenuous

c. redoubtable

d. vituline

(V-11) The _____ bellboy was intent on showing Jill all the features of the deluxe suite.

a. disingenuous

b. dilatory

c. officious

d. pugnacious

(V-12) She was _____ in her refusal to listen to our complaints.

a. officious

b. dilatory

c. trenchant

d. disingenuous

(V-13) The talks were _____ and Gates held little hope that the two companies would get together.

a. truculent

b. dilatory

c. conciliatory

d. craven

(V-14) Holmes plays a well-meaning but _____ lawyer who tries to make the grieving families sue for damages.

a. officious

b. sycophantic

c. redoubtable

d. dilatory

(V-15) The idea that he was _____, that he went about with a chip on his shoulder, that he loved fighting for the sake of fighting, was a mistake.

a. disingenuous

b. truculent

c. pugnacious

d. hidrotic

(V-16) Hoping to end the coldness that had grown between them, he wrote a _____ note.

a. craven

b. dilatory

c. truculent

d. sycophantic

(V-17) When Jack claimed he hadn't eaten the jelly doughnut, Jill took _____ look at his smeared face and laughed.

a. a spoffish

b. a redoubtable

c. a vituline

d. an officious

(V-18) Mahler's revolutionary music, abrasive personality and _____ writings about art and life divided the city into warring factions.

a. officious

b. trenchant

c. spoffish

d. pugnacious

(V-19) Donna thought it was _____ of Alex to retract his allegations instead of defending his position.

a. pugnacious

b. craven

c. disingenuous

d. conciliatory

(V-20) His voice has a _____ quality that is more appropriate at a funeral than a class reunion.

a. sycophantic

b. vituline

c. `pugnacious`

d. `hidrotic`

Definitions of obscure vocabulary words

- `hidrotic` — Pertaining to hidrosis, sweating, the production of perspiration.
- `spoffish` — Earnest and active in matters of no moment; bustling.
- `vituline` — Of, relating to, or resembling veal or a calf.

A.5 Phase 2: Instructions to Participants

Imagine that you work for the ABC Railroad. ABC train conductors take notes about the train routes and the arrival and departure times. During the past week, the conductors had to take notes by hand, on a company-provided illustrated pad. A friendly coworker calls you on the phone with questions about the train routes and the arrival and departure times. Your task is to answer the questions as honestly as possible. Some of the conductors have sloppy handwriting. At times, you may have trouble reading their notes.

A.6 Phase 2: Instructions to Annotators

We will ask you to listen to (and rate) a corpus containing 1276 short spoken utterances. The utterances were collected in the following manner: speakers were presented with an illustration of a train route and were asked to answer a question based on the information in the illustration. Speakers were told to answer the question aloud upon hearing a beep. The utterance recordings start immediately at the beep and end when the speaker finishes answering the question.

For each utterance, judge how certain you feel the speaker sounds. Use the following scale:

IMPORTANT: You are not asked to judge how sensible the sentence is, just judge how certain the speaker sounds. Further, you do not need to think about

the source of the certainty or uncertainty; make your judgments with respect to the whole utterance.

You may listen to each utterance as many times as you like. However, you should not go back and change previous ratings you have made.

The corpus is divided into 12 sections. We expect that each section will take 20-30 minutes. We will specify the order in which you should work through the sections. In each section, listen to the speech files in the order they are listed. You must complete each section in a single sitting, and do not do more than two sections without taking a break.

A.7 Phase 2: Mechanical Turk HIT Parameters

Required worker qualifications :

- HIT approval rate $\geq 95\%$
- Number of HITs approved ≥ 50
- Location is United States

HIT Instructions :

For each of the handwritten digits below, identify the digit using the drop-down menu. Even if you are unsure, select the digit that the image most closely resembles. We will compare your selections (for certain images) with the selections of other workers to ensure quality.

Number of images per HIT : 22 (including 2 control images)

Worker payment : \$0.05 per HIT